УДК 811.111

КОРПУСНАЯ ЛИНГВИСТИКА И ЕЕ ИСПОЛЬЗОВАНИЕ В КОМПЬЮТЕРИЗИРОВАННОМ ЯЗЫКОВОМ ОБУЧЕНИИ

О.В. Нагель

Аннотация. Рассматривается использование методов прикладной лингвистики в преподавании иностранного языка. Анализируется потенциал обучения на базе корпусного языкового материала. Представлен анализ таких методов корпусной лингвистики, как автоматизированное извлечение информации (information retrieval (IR)), обучение на основе данных (data-driven learning), текстовые поиски в крупномасштабных корпусах (конкордансы), методы обработки естественного языка (natural language processing (NLP)).
Ключевые слова: корпусная лингвистика, языковое обучение, конкорданс.

Одной из важнейших проблем, отмечаемых в процессе преподавания иностранных языков, является нехватка адекватных педагогических текстовых материалов и актуальных вокабуляров, в то время как ежедневная нагрузка преподавателей по текущей подготовке новых мотивирующих материалов для занятий продолжает оставаться стабильно высокой. В данной области значительную помощь может оказать привлечение методов корпусной лингвистики: ***автоматизированное извлечение информации***, ***обучение на основе данных***, текстовые поиски в крупномасштабных корпусах с использованием методов ***обработки естественного языка***.

Понятие «корпус текстов», на базе которого развивается корпусная лингвистика, все шире входит в научный оборот лингвистов. Под корпусом текстов обычно понимают «унифицированный, структурированный и размеченный массив языковых (речевых) данных в электронном виде, предназначенный для определенных филологических и, более широко, гуманитарных изысканий» [1. С. 52]. Корпус текстов может также рассматриваться «как сложно организованная онтология речевой деятельности, отражающая в себе все жанровое разнообразие представленного в нем рода словесности» [2. С. 18]. Как сложное словесное единство корпус включает в себя разнообразную информацию не только о составе и структуре своего речевого материала, но также и другие формализованные методы его представления (индексирование слов, морфологическая информация и т.д.). Следовательно, его также можно рассматривать как специальным образом построенную семиотическую систему [2. С. 21].

Корпусная лингвистика в современном ее понимании как наука, занимающаяся созданием и анализом текстовых корпусов, зародилась в США и Западной Европе в конце 1960-х гг. С ростом возможностей современных компьютерных технологий, с середины 1980-х гг. корпусная лингвистика получает бурное развитие, стали активно появляться корпус-

ные проекты различных масштабов на разных языках и для разнообразных целей.

Достижения в области корпусной лингвистики находят широкое применение в процессе преподавания языкознания. В ведущих вузах мира становится повседневной практикой использование корпусных данных в качестве эмпирической составляющей лекционных курсов, студенческих заданий и самостоятельных проектов. При этом оказывается, что корпусный подход оптимален для наглядного представления таких аспектов языка, как историческая, географическая и социальная вариация и изменения в языковой системе, параллельно давая живую возможность овладеть базовыми принципами корпусных методов лингвистического анализа [3].

Автоматизированное извлечение информации

Влияние современных технологий на обучающихся иностранному языку, возникновение новых технологических видов функционирования языка приводят к переосмыслению современного определения коммуникативной компетенции. Оперируя такой концептуализацией языка, *как контекстно-корректное использование различных регистров речи* [3], понятие коммуникативной компетенции нужно определять *как соответствующее естественным конвенциям* (*как грамматически, так и стилистически*) *использования языка в различных коммуникативных ситуациях, в том числе тех, в которые вовлечены технологии*. В этой связи неизбежным и необходимым является внедрение технологии во все аспекты прикладной лингвистики не как отдельного предмета, а как незаменимого исследовательского и обучающего инструмента и объекта критического анализа [3]. Существуют многочисленные экспериментально-обоснованные свидетельства того, что студенты, поощряемые к самостоятельному формированию собственного понимания черт изучаемого языка, овладевают языковыми компетенциями быстрее и эффективнее, чем те, кому вбиваются в голову правила (которые на поверку часто оказываются неадекватными реальному состоянию языка). Кроме того, опыт коллективного «открытия» языка в среде студенческой группы с помощью этих новых методик привносит ценнейший элемент обсуждения и совместных обобщений.

В последние 10–15 лет исследователи методик преподавания иностранных языков использовали обширные корпуса текстов для оценки реалий языка в его естественном состоянии. Эти корпуса текстов в значительной степени повлияли на повышение качественного уровня выпускаемых языковых пособий. Вместо традиционных прескриптивистских указаний, как надлежит правильно использовать язык, новые корпусные исследования описывают, эмпирически обоснованно анализируют то, что люди действительно говорят. Особого упоминания заслуживают новые словари, соз-

даваемые с использованием методик корпусной лингвистики, такие как Longman, Oxford, Collins, а также опыт критического переосмысления постулатов описательной грамматики английского языка (Longman Grammar of Spoken and Written English, опубликованной в 2000 г.).

Полученный опыт, как и общепринятая практика, говорит о том, что наиболее стабильные результаты получаются при пошаговом внедрении этих новых методик в процесс обучения для обеспечения эффективности и последовательной мотивации. Таким образом, на начальной стадии должны использоваться тщательно отобранные цитаты в раздаточном материале с хорошо проработанными инструкциями и заданиями, а на последующих этапах студенты смогут адекватно справляться с непредсказуемостью «живого» поиска конкордансов в Интернете и самостоятельно формулировать исследовательские задачи. Наиболее интересными в данном аспекте являются критические аналитические работы по современной английской грамматике, полученные в области естественного разговорного языка, который на поверку обнаруживает гораздо более обширные расхождения со стандартным письменным языком, чем на это когда-либо указывалось в учебно-методических пособиях. Теперь же, будучи выявленными при помощи корпусных методик, эти особенности могут быть учтены в процессе преподавания и при разработке современных учебно-методических комплексов.

Обучение на основе данных (data-driven learning)

В последнее десятилетие зародилось еще одно новое и чрезвычайно перспективное направление в организации процесса обучения иностранным языкам, при котором студент имеет возможность прибегать к использованию «сырых» языковых данных напрямую из корпуса. Это направление получило название *обучение на основе данных*, или data-driven learning (DDL). Оно основано на солидном эмпирическом доказательстве того, что студенты могут гораздо более эффективно осваивать язык, когда в процессе обучения поощряется использование модели *наблюдай – предполагай – экспериментируй* (observe – hypothesize – experiment model), т.е. когда они имеют возможность делать собственные выводы относительно значений слов, фраз, грамматических правил на основе аутентичного языкового материала. Этот индуктивный метод дополняет более распространенный дедуктивный подход, известный также как *слушай – практикуйся – говори*, при котором студенты получают знание о правилах и определениях из объяснений инструктора и справочной литературы.

Процесс не обязательно ограничен терминалом компьютера. Результаты корпусных поисков (конкордансов) в распечатанном виде могут быть легко инкорпорированы в раздаточный материал, методические пособия и т.п. и использованы в процессе традиционного преподавания на уроке.

Кроме этого, достаточно распространенным является формирование специализированных корпусов текстов на жестком диске. Современные средства позволяют быстро сформировать весьма обширный (несколько десятков миллионов слов) корпус текстов практически по любой тематике, и сделать это может каждый, кто владеет основными навыками работы с персональным компьютером и Интернетом.

Методы обработки естественного языка

В *компьютеризированном языковом обучении* (КЯО) (computer-assisted language learning (CALL)) на данный момент идут активные разработки в области так называемых *языковых технологий* с использованием методов *обработки естественного языка* (natural language processing (NLP)) для создания «умных» методик. В таких инновационных методиках используются лингвистический анализ и моделирование реакций обучающегося для анализа и адекватной оценки так называемого *языкового материала в свободной форме* (free-form linguistic input) со стороны обучающегося, другими словами – ответы на открытые вопросы и даже оценка свободно оформленных эссе.

Обучение языку при помощи компьютерных технологий отходит от основных, традиционных, чаще всего деконтекстуализованных способов подачи материала, и фокусирует внимание на тех видах деятельности, которые стимулируют или, даже можно сказать, требуют элемент творчества. Карол Чапэл в статье «Практическое применении высоких технологий в преподавании» отмечает, что, например, «в рамках курса грамматического анализа мы вовлекаем студентов в анализ компьютерного корпуса, что, во-первых, влияет на их восприятие самого анализа, во-вторых, на их способность самостоятельно проводить подобный анализ и, в-третьих, на то, как они сами будут преподавать грамматику» [4. С. 6].

Содержание методических материалов и практика преподавания иностранных языков и языкознания как у нас в стране, так и повсеместно имеют тенденцию отражать то разделение, которое существует на данный момент между эмпирическим и рационалистским подходами в гуманитарных науках, в частности в языкознании. Многие учебники изобилуют искусственными примерами, в то время как грамматические и стилистические описания основываются в большей мере на интуиции их составителей или на вторичных источниках. Однако существует небольшое число учебных пособий, которые основаны на эксплицитно эмпирическом подходе и используют примеры и описания, почерпнутые из корпусов реально используемых языковых средств.

Естественные языковые средства (Naturally-occurring Language) чрезвычайно важны в процессе обучения иностранным языкам, т.к. предоставляют студентам возможность иметь дело с теми предложениями,

которые они встретят в реальной ситуации общения на иностранном языке. Студенты, которые обучаются на основе консервативных учебных материалов с традиционными описаниями письменного синтаксиса типа *Mary puts her book on the table*, порой не готовы воспринимать и корректно анализировать естественную речь, изобилующую сложными предложениями типа *The government has welcomed a report by an Australian royal commission on the effects of Britain's atomic bomb testing programme in the Australian desert in the fifties and early sixties* (из Корпуса Разговорного Английского = Corpus of Spoken English) [5].

Кроме прямого применения в процессе преподавания иностранного языка на основе естественного эмпирического подхода, корпус как метод может быть использован для критической оценки существующих методических материалов. Так, Кэннеди, Холмс, Миндт анализировали освещение различных аспектов грамматики английского языка в существующих традиционных пособиях, используя методику сравнительного анализа соответствующих конструкций и вокабуляра в учебниках и в корпусе стандартного английского. В ходе большинства таких исследований было обнаружено, что существуют значительные расхождения между тем, что предписывается учебниками и тем, как язык действительно используется носителями, о чем свидетельствует корпус разговорного языка. Порой в некоторых учебных пособиях на передний план ставятся отдельные аспекты использования языка и его стилистических особенностей, которые оказываются периферийными и менее типичными, в то время как более центральные игнорируются. Общим выводом этих исследований является то, что традиционные прескриптивистские учебные материалы, не основанные на эмпирических методах отбора и анализа языкового материала, неадекватны реальному естественному состоянию языка и реалиям его типичного применения, а также то, что методы корпусной лингвистики должны быть обязательны при разработке и оценке эффективности учебных материалов и методических пособий с тем, чтобы наиболее распространенные употребления получали приоритетное внимание, а периферийные употребления занимали соответствующее им место [6].

Таким образом, роль корпусного подхода в такой области, как компьютеризированное обучение иностранным языкам центральна. Последние исследования Университета Ланкастера, посвященные программному обеспечению для обучения студентов младших курсов грамматике и основам грамматического анализа, показали, что такие программы, как Cytor и аналогичные им, создаваемые достаточно легко на основе корпуса текстов аннотированного или по частям речи или по грамматическим/синтаксическим ролям, чрезвычайно эффективны и обеспечивают нужную степень интерактивности наряду с автономностью [3]. Получая задание грамматического разбора текста со скрытой аннотацией, студенты самостоятельно разбирают предложения, имея возможность запросить у программы помощь в виде списка обозначений информации о частотно-

сти употребления той или иной лексической единицы или частотности совместного употребления примеров (коллокации).

Приведем частный пример возможностей использования корпуса текстов и таких методов, как поиск конкордансов в преподавании ИЯ. Задание для студентов будет звучать следующим образом:

ВОПРОС: *Какова разница между словами remember и remind? Правильно ли следующее?*

Do you remember me? I used to sit at the back of your class (*correct*).

Can you remember me? I used to sit at the back of your class. (*Is this wrong?*)

The flowers reminded him his garden. (*Is this wrong?*)

Можете ли вы дать мне набор предложений (*около 20*)*, где бы я мог заполнить пробелы, используя* «*remember*» *или* «*remind*».

В результате студенты в режиме конкорданса получают набор предложений с нужным словом и в процессе контекстного анализа выявляют семантическую разницу [7].

В ходе экспериментов, проведенных МакЭнри, Бэйкером, Уилсоном с целью определения эффективности этой методики, была установлена степень усвоения знания частей речи среди студентов, обучавшихся по новой – корпусной методике, и в контрольной группе студентов, обучавшихся по традиционной – лекционной методике [8]. В целом студенты, обучавшиеся с помощью корпусного программного обеспечения, последовательно демонстрировали более высокие показатели, нежели студенты контрольной группы. Рассматривая вопросы, с которыми ежедневно сталкиваются студенты на различных стадиях овладения иностранным языком, можно отметить, что методы корпусной лингвистики, например поиск конкордансов, оптимально приспособлены для того, чтобы обеспечить интересный и самостоятельный поиск ответов, в процессе которого студент имеет возможность как получить искомые сведения по частным вопросам изучаемого языка, так и попутно приобрести представление о реальном естественном состоянии языка, его историческом, географическом и социальном варьировании, регистрах речи, жанровом разнообразии.

Важно отметить, что корпуса текстов, как и современная корпусная лингвистика, как научная методология и отрасль языкознания, очень молоды и развиты неравномерно. Для некоторых языков, таких как английский, немецкий, финский или японский, созданы обширные и репрезентативные аннотированные корпуса, в то время как для других языков, включая и русский, процесс создания полноценных корпусов, соответствующих основным требованиям, переживает период становления. Ситуация с мультиязычными (параллельными или сравнительными) корпусами еще более сложная, что значительно ограничивает невероятный потенциал корпусных методов в области лингвистических исследований, языковых технологий и преподавании иностранных языков.

Из осуществляемых в данный момент различных моно- и мультиязычных корпусных проектов, большинство существует на базе университетов. Например, на базе Томского политехнического университета успешно была реализована *Комплексная программа развития* на 2004–2005 гг. «Корпус текстов как разновидность языковых ресурсов», разработанная доцентами Н.Н. Шаламовой и А.Ю. Фильченко [9]. В ходе этой работы в 2004 г. были выполнены три выпускные квалификационные работы в рамках единого тематического направления «Корпусов инженерных текстов на английском языке» для последующего его использования в практике преподавания английского языка и научно-исследовательской работе в ТПУ.

Корпусные методы зарекомендовали себя в мировой практике лингвистических исследований и преподавании иностранных языков, включая профильно-ориентированное, как высокоэффективные инновационные дополнения к традиционным образовательным технологиям. Эти методы сочетают в себе такие аспекты, как междисциплинарность, эмпирическая адекватность, аутентичность, гибкость и адаптация к конкретным задачам и целевым группам, возможность самостоятельной работы студента, применение метода «открытия» в обучении.

Литература

1. ***Захаров В.П.*** Поисковые системы Интернета как инструмент лингвистических исследований // Русский язык в Интернете. Казань, 2003.

2. ***Рыков В.В.*** Корпус текстов как новый тип словесного единства // Труды Междунар. семинара «Диалог-2003». М.: Наука, 2003. С. 15–23.

3. ***Information*** and Communications Technology for Language Teachers. Введение в прикладное значение корпуса. Режим доступа: http://www.ict4lt.org/en/en_mod2-4.htm

4. ***Carol A.*** Chapelle // Essential teacher. 2003. Vol. 9. P. 5–11.

5. ***British*** National Corpus. Режим доступа: http://sara.natcorp.ox.ac.uk/lookup.html

6. ***National*** Foreign Language Resource Center. Режим доступа: http://www.nflrc.hawaii.edu/project/2006EmoreInfo.html

7. ***Простой*** пример конкорданса на сайте корпуса Cobuild. Режим доступа: http://titania.cobuild.collins.co.uk/form.html

8. ***McEnery T., Wilson A.*** Corpus Linguistics. Edinburgh, 1997.

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

CORPUS LINGUISTICS AND ITS USE IN COMPUTER-BASED LANGUAGE TEACHING

Nagel O.V.

Summary. This paper is concerned with certain methods of applied linguistics used in language teaching. Such methods of corpus linguistics as information retrieval (IR), data-driven learning, natural language processing (NLP) and their application in foreign language acquisition are under consideration.

Key words: corpus linguistics, language teaching, concordance.