

К. Р. Пиотровская, В. Р. Нымм,

Российский государственный педагогический университет им. А. И. Герцена, Санкт-Петербург

МНОГОЦЕЛЕВАЯ СИСТЕМА ЭЛЕКТРОННОЙ ПОДДЕРЖКИ ОБУЧЕНИЯ ИНОСТРАННОМУ ЯЗЫКУ И ТЕКСТ-МАЙНИНГОВЫЕ ОТКРЫТЫЕ РЕСУРСЫ

Аннотация

В статье описана многоцелевая система электронной поддержки обучения иностранному языку на базе вычислительного эксперимента, которая позволяет осуществлять подготовку содержания обучения, языковое компьютерное обучение и тестирование, реализуя четыре основных типа письменных упражнений типа «стимул — реакция». Система работает при поддержке текст-майнинговых открытых сервисов.

Ключевые слова: лингво-статистический подход, вычислительный эксперимент, адаптивное обучение, открытые текст-майнинговые веб-сервисы.

Стремительный темп информационного обмена, ведущий к изменчивости лексико-грамматического состава профессионально-ориентированных языков, требует постоянной, оперативно организованной оптимизации процесса формирования иноязычной профессиональной компетенции у студентов, обучающихся по самым разным направлениям науки и техники. Появление новых надежных открытых сервисов, оперативно осуществляющих автоматизированное статистическое исследование текста, и многоцелевых лингводидактических средств электронной поддержки позволяет оперативно формировать базы языковых упражнений для отработки терминологии, грамматических особенностей и фразеологических оборотов английского языка в технической и деловой сферах различной жанрово-стилевой направленности: научных статьях, инструкциях для различной аппаратуры, докладах для международных журналов и конференций, рабочих отчетах.

Квантитативные методы могут эффективно применяться при отборе содержания обучения и его организации в учебные терминологические словари, списки терминов и устойчивых словосочетаний. Статистиче-

ские методы позволяют выделять лингвистические единицы, обладающие наибольшим удельным информативным весом в процессе коммуникации, в некий минимум, который достаточен для понимания значительной части текста определенной тематики. Терминологические единицы, отобранные статистическим путем, далее могут организовываться в парадигмы исходного языка предметной области знаний путем автоматизированного выделения устойчивых словосочетаний и построения конкордансов. Полученный таким образом базовый язык представляет собой не просто статистическое, но и структурное построение. Хотя объективные методы отбора учебных терминологических единиц (например, статистико-вероятностный подход) пропагандируются начиная с шестидесятых годов XX века в работах П. М. Алексеева, Л. П. Герман-Прозоровой, А. К. Головачевой, В. П. Григорьева, Р. Г. Пиотровского, О. П. Щепетовой и др. [1, 6], однако до сих пор отбор терминологических единиц для изучения в рамках профессионально-ориентированного языка производится подчас субъективно, исключительно с опорой на личный опыт, компетентность и интуицию преподавателей.

Контактная информация

Пиотровская Ксения Раймондовна, доктор пед. наук, канд. тех. наук, доцент, профессор кафедры методики обучения математике и информатике Российского государственного педагогического университета им. А. И. Герцена, Санкт-Петербург; *адрес:* 191186, г. Санкт-Петербург, наб. реки Мойки, д. 48; *телефон:* (812) 312-44-92; *e-mail:* krp62@mail.ru

Нымм Волдемар Рихардович, канд. пед. наук, доцент, доцент кафедры образовательных технологий в филологии Российского государственного педагогического университета им. А. И. Герцена, Санкт-Петербург; *адрес:* 191186, г. Санкт-Петербург, наб. реки Мойки, д. 48; *телефон:* (812) 312-44-92; *e-mail:* jetnomm@gmail.com

X. R. Piotrowska, V. R. Nymm,

The Herzen State Pedagogical University of Russia, Saint Petersburg

MULTI-PURPOSE COMPUTER ASSISTED LANGUAGE LEARNING PROGRAM AND OPENED TEXT-MINING RESOURCES

Abstract

An approach to solve the CALL-problem in the frameworks of the computing experiment is proposed. A special attention is drawn to the development of the methodological, informational and organizational components of the system. The system is supported by the text-mining opened services.

Keywords: computing experiment, text-mining, computer-assisted language learning program.

Сегодня статистико-вероятностный подход переживает некоторый ренессанс [7, 10, 12], а большинство западных исследователей трактуют наличие лингво-статистических элементов в системах как элементы текст-майнинга [14, 16]. Для поддержки подхода международная группа исследователей из Японии, Великобритании и Новой Зеландии под руководством профессора Л. Антони, директора Центра обучения профессиональному английскому языку в Школе науки и техники Университета Васэда (Япония), создала ряд открытых сервисов под общим названием AntConc [13, 15]. Семейство этих программ используется в корпусных исследованиях и, как нам представляется, может оказаться полезным инструментом при отборе содержания обучения для лингводидактических комплексов.

В течение ряда лет рабочая группа Российского государственного педагогического университета им. А. И. Герцена развивает **концепцию создания, функционирования и развития CALL-технологии** (названной по известной в лингводидактике аббревиатуре от Computer Assisted Language Learning), ориентированной как на поддержку обучения иностранным языкам, так и на выполнение научных и научно-педагогических исследований по проблеме компьютерного обучения [8]. В этом контексте создание на базе открытых сервисов AntConc технологической цепочки по количественному отбору содержания баз обучающей информации и включение ее в состав CALL-технологии, с одной стороны, расширяет границы концепции и придает ей большую целостность, с другой — создает новую область для исследований и разработок студентов в рамках выпускных квалификационных работ и диссертаций разного уровня.

Рассмотрим основные составляющие CALL-технологии в ее расширенном виде.

Постановка задач обучения

Как и в работе [11], посвященной исследованию бесконтекстного обучения лексике иностранного языка, мы рассматриваем две задачи. Первая определяет желаемое состояние обучаемого на момент завершения процесса обучения и выражается в форме неравенства $Q(p) < q$, где $Q(p)$ — функционал качества обучения, $p(t) = (p_1(t), p_2(t), \dots, p_n(t))$ — вектор параметров состояния обучаемого, t — время, а q — пороговое значение функционала. Вторая задача является экстремальной и состоит в минимизации времени достижения желаемого состояния обучаемого.

Учитывая достаточно опосредованную связь между указанным выше неравенством и временем, за которое это условие должно быть достигнуто, поиск решения подобных задач, как правило, переводится на язык стратегий. «В понятии минимального времени нет глубокого смысла, тогда как в понятии оптимальной стратегии такой смысл есть. Здесь интуиция и опыт действительно могут нам помочь. На самом деле вы знаете о процессе значительно больше, чем можете выразить явно в математической форме, и это помогает вам угадать приближенную стратегию» [3].

Следование этой установке переводит решение задачи в область поиска эвристического построения,

обеспечивающего достаточно быструю сходимость функционала $Q(p)$ к пороговому значению. При этом сам поисковый процесс, как правило, реализуется в рамках схемы, называемой вычислительным экспериментом.

Вычислительный эксперимент и условия его реализации

В методологическом плане под **вычислительным экспериментом** здесь понимается «циклический процесс, реализующий эволюционную стратегию построения модели исследуемого объекта или процесса, каждую итерацию которой составляют три последовательно реализуемые стадии: анализ, синтез и оценка, а проверка адекватности модели, синтезированной в ходе каждой очередной итерации, осуществляется путем сравнения результатов натурных экспериментов с результатами, полученными с помощью компьютера путем вычислений по модели» [4].

Для осуществления вычислительного эксперимента на филологическом факультете РГПУ им. А. И. Герцена была разработана **многоцелевая программа компьютерного обучения и тестирования CALL-program**, предназначенная для обучения языку и лингводидактических исследований студентов и аспирантов.

Учитывая тот факт, что именно алгоритм управления процессом обучения воплощает в себе решение поставленных ранее задач обучения, структура программы изначально проектировалась так, чтобы обеспечить «безболезненную» (т. е. без сколько-нибудь существенных перестроек программы) замену одних алгоритмов управления другими.

В совокупности со встроенными в программу средствами сбора информации о ходе процессов обучения реализация этого условия обеспечивает механизм параллельного развития как самой программы, так и проводимых с ее помощью исследований:

- каждое организованное (т. е. проводимое либо в рамках выполнения студентами заданий учебного плана, либо в исследовательских целях) использование программы рассматривается как элемент массового эксперимента;
- получаемые в ходе каждой итерации технологического цикла вычислительного эксперимента данные о ходе обучения обрабатываются и преобразуются в новые знания о процессах компьютерного обучения иностранным языкам;
- на основе полученных знаний ставятся новые задачи исследования, создаются новые и корректируются существующие алгоритмы управления процессом обучения, создаются новые базы обучающей информации и корректируются состав и объем существующих;
- производится замена одних алгоритмов управления другими и осуществляется переход к следующей итерации технологического цикла.

Таким образом, в любой момент времени программа работает на основе алгоритма, соответствующего существующим знаниям и представлениям о процессах компьютерного обучения.

Алгоритм управления процессом обучения

Предмет одного из направлений исследования и разработки по проблеме компьютерного обучения языкам — алгоритмы, построенные на базе одноэлементной модели обучения парным ассоциациям, которую иногда называют «всё или ничего» [2]. В основе этой модели лежит утверждение о том, что ассоциация, связывающая стимул и реакцию, не может формироваться частично или постепенно. Предполагается, что, однажды возникнув, ассоциация сохраняется в течение длительного времени и обучаемый будет давать верный ответ при каждом новом предъявлении ему стимула. Для одноэлементной модели состояние обучаемого в любой момент времени описывается вектором $p(t) = (p_1(t), p_2(t), \dots, p_n(t))$, каждая координата которого имеет значение 0, если по i -му элементу ассоциация не сформирована, и значение 1 — в противном случае.

Полное описание алгоритма приведено в статье [9], здесь же мы отметим лишь его существенные характеристики и результаты тестирования разработанного алгоритма.

Для задачи обучения лексики иностранного языка в качестве критерия завершения обучения по каждому отдельному упражнению рассматривается событие, состоящее в том, что обучаемый в ходе одного из сеансов дает верный ответ с первого же в текущем сеансе предъявления стимула этого упражнения. Если считать, что в день реализуется не более одного сеанса, то это означает, что верный ответ с первого раза был дан не менее чем через 24 часа после предыдущего предъявления этого стимула. Обучение в целом продолжается до тех пор, пока все упражнения, представленные в базе обучающей информации, не окажутся помеченными как элементы, по которым обучение завершено, т. е. когда значения всех координат вектора p (вектора параметров состояния обучаемого) станут равными 1.

По окончании процесса обучения обучаемый допускается к контрольному тестированию. Тестирование может осуществляться либо по всему представленному материалу, либо по его достаточно представительной случайной выборке.

Аналитика процесса обучения данной программы включает:

- количество сеансов, которое потребовалось обучаемому до завершения обучения по каждому упражнению;
- число предъявлений стимула упражнения в ходе этих сеансов;
- значения параметров, определяющих динамику процесса;
- результаты контрольного тестирования.

Эффективность алгоритма оценивалась по результатам эксперимента, в котором принимала участие группа из 50 студентов филологического факультета. Материал изучения включал 100 языковых клише (устойчивых словосочетаний, вводных конструкций и т. п.) по общенаучной тематике. Процесс обучения осуществлялся в ходе последовательных сеансов. Интервал времени между сеансами у разных студентов колебался от 24 до 72 часов.

На основе обработки данных были получены следующие результаты. Среднее число предъявлений стимула (по группе в целом), которое потребовалось для завершения обучения по одному упражнению, оказалось равным 4,64. Разброс этого показателя для отдельных обучаемых достаточно велик: от 2,72 до 7,81, стандартное отклонение — 1,57.

Среднее число сеансов, в ходе которых эти предъявления реализовывались, оказалось достаточно малым — всего 1,91. Это означает, что «выучивание» правильного ответа в среднем осуществлялось примерно за два сеанса так, что на третьем сеансе обучаемый мог воспроизвести правильный ответ с первого раза. Разброс значений этого показателя также достаточно мал: от 1,51 до 2,37, стандартное отклонение — 0,32.

Непосредственные результаты обучения оценивались по данным контрольного тестирования. Несмотря на то что процедура тестирования была достаточно утомительной (она проводилась по всему материалу обучения, а не по выборке из него, как это делается обычно), результаты обучения оказались более чем оптимистичными. Доля правильных ответов для разных обучаемых оказалась в диапазоне от 0,61 до 0,9, среднее значение — 0,779, стандартное отклонение — 0,0832.

Технология создания баз обучающей и тестовой информации

Для оптимизации процесса пополнения баз обучения и тестирования мы обратились к некоторым из открытых текст-майнинговых сервисов, созданных исследовательской группой из японского Университета Васэда под руководством профессора Л. Антони.

В семейство программ AntConc входят:

- AntConc — конкордансер, который занимает центральное место в описываемом семействе и позволяет исследовать частотный состав текста, выявлять коллокации — слова, с которыми анализируемое слово (часто) встречается, строить конкордансы;
- AntPConc — построитель параллельных текстов для последующего анализа текста и построения конкордансов;
- AntMover — анализатор семантической структуры текста;
- AntWordProfiler — инструмент для профилирования словаря и проверки сложности текста;
- ProtAnt — инструмент для поиска файла прототипа;
- TagAnt — инструмент для теггирования по частям речи;
- а также ряд сервисов для перекодировки символов, спеллеров и подготовки текстов к последующей обработке (AntFileConverter, EndCodeAnt, VariAnt).

Особняком стоят различные сервисы в помощь при работе с иероглифическими текстами. Проект находится в постоянном развитии, появляются новые улучшенные версии уже надежно работающих сервисов и разрабатываются новые.

Нам представляется возможным организовать следующую технологическую цепочку для оптимизации отбора учебного контента и последующего формирования упражнений в CALL-программе (см. рис.):

- предварительно студент должен перевести заранее отобранные тексты в нужный формат, используя программы AntFileConverter, EndCodeAnt и VariAnt;
- далее с помощью программы AntConc построить частотный словарь, выделить ключевые слова и выявить устойчивые словосочетания с ключевыми словами;
- объем исследуемых текстов далее наращивается путем использования программы ProtAnt по введенным ключевым словам, которые выявлены на предварительном этапе или по эталонному тексту; эта программа осуществляет отбор

текстов, наиболее близких по лексическому составу к поставленной учебной задаче;

- после этого необходимо вернуться снова к работе по лингво-статистическому исследованию в программе AntConc, но уже расширенного состава учебных текстов, а также к профилированию полученного словаря и разметке текстов;
- для удобства формирования учебных упражнений с помощью программы AntMover производится фрагментация и структурно-семантическая разметка текстов по предложениям с сохранением каждого предложения в отдельный файл;
- далее программой ProtAnt снова обрабатываются уже файлы отдельных предложений и выявляются те из них, которые по лексическому или грамматическому составу подходят для генерации упражнений в CALL-программу.



Рис. Технология отбора содержания обучения на базе семейства программ AntConc

Реализация вычислительного эксперимента в условиях вуза

В организационном плане реализация вычислительного эксперимента в условиях вуза опирается на метод *learning by doing* («обучение в деле») и рассматривается как часть высокотехнологичной образовательной среды, в рамках которой осуществляется интеграция учебной и созидательной деятельности студентов в области информационных технологий, направленная на создание компонентов информационного и методического обеспечения систем языкового компьютерного обучения, с одной стороны, с другой — на то, чтобы в процессе этой деятельности студенты научились:

- использовать программу языкового компьютерного обучения и тестирования для выполнения заданий учебного плана;
- создавать (на основе существующего программного обеспечения) собственные системы компьютерного обучения;
- оперативно осуществлять отбор и пополнение содержания обучения с использованием текстмайнинговых сервисов;
- выполнять на их базе научные и научно-педагогические исследования;
- обрабатывать полученные данные о ходе обучения, интерпретировать результаты.

Такой подход, с одной стороны, придает обучению четко выраженный прикладной характер, позволяет продемонстрировать полную цепочку операций моделирования на базе ИТ-процессов, относящихся к области основной специализации студентов, с другой — дает возможность использовать созидательную практическую деятельность студентов для реализации двух важных функций вычислительного эксперимента:

- создания, тестирования и корректировки электронных баз обучающей (БОИ) и тестовой (БТИ) информации по иностранным языкам;
- выполнения лабораторных экспериментов, в роли участников которых выступают сами студенты.

Первая из перечисленных функций позволяет избежать достаточно типичных ситуаций, когда при полной готовности к выполнению эксперимента практически невозможно найти участников (тем более мотивированных) для его реализации. Еще более очевидна важность второй функции. Как показывают результаты внедрения подхода, интеллектуальный ресурс, который студенты составляют в совокупности, и временной ресурс, выделяемый учебным планом на самостоятельную работу студентов, достаточны для помощи при создании и актуализации БОИ и БТИ.

Заключение

Дальнейшие шаги по развитию программного обеспечения CALL-технологии связаны как с расширением работ по исследованию алгоритмов управления процессами обучения, так и с увеличением технологичности разрабатываемых программных продуктов. В настоящее время ведутся совместные

работы филологического факультета, лаборатории инженерной лингвистики и Института компьютерных наук и технологического образования РГПУ им. А. И. Герцена по созданию сетевой версии базовой программы. В части развития информационного обеспечения на базе создания профессионально-ориентированных лингводидактических комплексов ведется отработка методики количественного отбора их содержания.

Литература

1. Алексеев П. М., Герман-Прозорова Л. П., Пиотровский Р. Г., Щепетова О. П. Основы статистической оптимизации преподавания иностранных языков // Статистика речи и автоматический анализ текста. Л.: Изд. ЛГПИ, 1974.
2. Аткинсон Р., Бауэр Г., Кротерс Э. Введение в математическую теорию обучения. М.: Мир, 1969.
3. Беллман Р. Может ли машина мыслить? Проблемы современной кибернетики. М.: Знание, 1975.
4. Беляева Л. Н., Джепа Т. Л., Зак Г. Н., Камшилова О. Н., Нымм В. Р., Разумова В. В. Автоматизированное рабочее место филолога в структуре образовательного пространства вуза: монография. СПб.: Книжный дом, 2013.
5. Буш Р., Мостеллер Ф. Сравнение восьми моделей // Математические методы в социальных науках. М.: Физматгиз, 1973.
6. Головачева А. К. Частотный курс ускоренного обучения английскому языку по профилю радиоэлектроники Л.: Изд-во ЛГУ, 1978.
7. Лаздинь Т. А. Основы статистической оптимизации преподавания иностранных языков // Вестник Санкт-Петербургского университета. Серия 9. 2009. Вып. 3.
8. Нымм В. Р. Эволюционный подход к созданию и развитию CALL-технологий на базе вычислительного эксперимента // Известия Российского государственного педагогического университета им. А. И. Герцена. 2015. № 173.
9. Нымм В. Р., Пиотровская К. Р., Быстрая В. А. CALL-технологии: управление процессом обучения языку // Высокотехнологичная образовательная среда: сборник статей международной научно-практической конференции. СПб.: Книжный дом, 2015.
10. Пиотровская К. Р. Основы количественной лингвистики и компьютерное обучение языкам // Концепции университетского образования в новом тысячелетии: материалы международной научно-практической конференции. СПб.: Изд-во РГПУ им. А. И. Герцена, 2002.
11. Растринин Л. А., Эренштейн М. Х. Адаптивное обучение с моделью обучаемого. Рига: Зинатне, 1988.
12. Сергеевнина В. М. Из опыта составления учебных терминологических словарей для студентов неязыковых вузов // Вестник Нижегородского университета. 2011.
13. Anthony L., Baker P. ProtAnt: A Freeware Tool for Automated Prototypical Text Detection // Proceedings of Corpus Linguistics. Lancaster, UK: Lancaster University, 2015.
14. Anthony L., Bowen M. The Language of Mathematics: A Corpus-based Analysis of Research Article Writing in a Neglected Field // Asian ESP Journal. 2013. 9 (2).
15. Nation P., Anthony L. Measuring vocabulary size // Handbook of Research in Second Language Teaching and Learning, Vol. III. New York: Routledge, 2016.
16. Scrivner O., Davis J. Topic Modeling of Scholarly Articles: Interactive Text Mining Suite // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2016, Moscow, June 1–4, 2016.