ИЗДАТЕЛЬСТВО ГРАМОТА

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2024. Том 17. Выпуск 3 | 2024. Volume 17. Issue 3
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Генерация файлов формата Moodle XML на основе данных сбалансированного лингвистического корпуса

Горожанов А. И.

Аннотация. Цель предлагаемого исследования – разработать алгоритм генерации образовательного (лингводидактического) контента для популярных систем управления обучением на основе корпусных данных. Научная новизна заключается в том, что впервые в оригинальную авторскую концепцию метода генерации сбалансированного лингвистического корпуса добавляется лингводидактический компонент, реализуемый на общих принципах максимальной автоматизации и универсальности применения. В ходе работы были изучены структуры резервных файлов LMS Moodle, разработан программный модуль преобразования данных авторского сбалансированного лингвистического корпуса с морфологической разметкой в файлы формата XML, которые были импортированы и протестированы на предмет целостности и правильности с точки зрения решения поставленной лингводидактической задачи. В результате доказывается, что используемый лингвистический корпус обладает высоким лингводидактическим потенциалом, созданный алгоритм показывает хороший результат, в данном случае – для генерации тестовых или тренировочных заданий по немецкому языку, и может быть расширен на материале других языков, поддержка которых предусмотрена программным комплексом – генератором корпуса.

EN

Generation of XML Moodle format files based on data from a balanced linguistic corpus

Gorozhanov A. I.

Abstract. The aim of the proposed research is to develop an algorithm for generating educational (linguodidactic) content for popular learning management systems based on corpus data. The scientific novelty lies in the fact that for the first time, a linguodidactic component is added to the original author's concept of the method for generating a balanced linguistic corpus, implemented on the general principles of maximum automation and universality of application. During the research, the structures of LMS Moodle backup files were studied, a software module was developed for converting data from the author's balanced linguistic corpus with morphological markup into XML files, which were imported and tested for integrity and correctness in terms of solving the posed linguodidactic task. As a result, it was proved that the used linguistic corpus has a high linguodidactic potential, the created algorithm shows good results, in this case, for generating test or training tasks in German, and can be expanded to material from other languages, the support of which is provided by the software tool – the corpus generator.

Введение

В наши дни в рамках образовательного процесса в вузах используется большое количество формализованных «учебных» данных. К ним можно отнести, например, компоненты систем управления обучением различного уровня: от целых онлайн-курсов до отдельных упражнений. Модульный характер этих систем, к которым также относится популярная во всём мире *LMS Moodle*, позволяет резервировать отдельные учебные элементы (например, записи глоссария или тестовые задания) в виде организованных по особым правилам файлов. Это дает возможность хранить, группировать и легко переносить данные из одной инсталляции системы в другую.

Одним из популярных форматов резервных файлов является формат *XML* (*eXtensible Markup Language*), который, кроме прочего, очень удобен для программной обработки и широко применяется в России для структурирования данных (Шалаев, 2019).

Такие файлы резервных копий, как правило, генерируются внутри системы управления обучением с помощью встроенных опций, т. е. для их получения необходимо вначале создать тот или иной компонент

Научная статья (original research article) | https://doi.org/10.30853/phil20240131

системы «обычным» способом, предусматривающим большую долю ручного труда (копирование, вставку, форматирование, заполнение метаданных и пр.).

В настоящее время технологии обработки естественного языка позволяют идти и обратным путём: генерировать резервные копии из текстовых массивов, а затем разворачивать их в виде определённых учебных модулей в онлайн-курсах систем управления обучением.

В целом идея автоматической генерации образовательного контента не может считаться чем-то совершенно новым и уникальным, о чем свидетельствуют исследования последних лет (Лаптев, Ларченкова, Шубина, 2023; Личаргин, Бачурина, 2021). Для создания электронных учебных материалов или для иной поддержки образовательного процесса применяются нейросети (Донина, 2023; Евстигнеев, 2023) и корпусные технологии (Котюрова, 2023; Сысоев, Клочихин, 2022).

Безусловно, большое количество публикаций свидетельствует о высокой актуальности темы, однако считать её исчерпанной не представляется возможным, так как многие вопросы остаются нерешёнными. Среди них выделим проблему разработки отечественного программного обеспечения и проблему учёта при разработке программных решений принципов методики обучения (в нашем случае – иностранным языкам).

В настоящем исследовании мы ставим перед собой следующие задачи:

- провести анализ структуры резервных файлов *LMS Moodle* для определения максимально универсального и удобного для автоматической обработки формата;
- оценить возможность использования авторского сбалансированного лингвистического корпуса с морфологической разметкой для автоматической генерации образовательного (лингводидактического) контента для *LMS Moodle*;
- провести экспериментальную генерацию электронных учебных материалов для *LMS Moodle*, предназначенных для обучения иностранному языку.

При решении первой задачи используется метод анализа, изучаются структуры резервных файлов на примере *LMS Moodle*. Вторая задача также требует привлечения метода анализа возможностей сбалансированного лингвистического корпуса с последующим сопоставительным анализом совместимости его внутренней структуры с характеристиками резервных файлов системы управления обучением. Третья задача решается в ходе эксперимента по разработке и применению программных модулей на языке программирования *Python* и импортированию полученных резервных файлов в *LMS Moodle*.

Лингвистическим материалом исследования являются сбалансированные корпусы оригинальных текстов романов Э. М. Ремарка «На Западном фронте без перемен» и «Возлюби ближнего своего» объемом 69914 и 137926 токенов соответственно. Корпусы были составлены с помощью авторского программного комплекса (Свидетельство о государственной регистрации программы для ЭВМ № 2023683209 Российская Федерация. «Генератор сбалансированного лингвистического корпуса и корпусный менеджер»: № 2023682269: заявл. 25.10.2023: опубл. 03.11.2023 / А. И. Горожанов; заявитель – федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный лингвистический университет»), который можно обозначить как один из инструментов исследования.

В качестве технологического объекта (в нашей терминологии – узла институциональной обучающей виртуальной среды), в рамках которого применяется лингвистический материал, нами используется инсталляция *LMS Moodle* версии 3.0+.

Теоретической базой исследования послужили труды Д. В. Степановой (2023) и А. И. Горожанова (Горожанов, Степанова, 2022), Р. К. Потаповой (2021), А. В. Зубова (2006) в области прикладной лингвистики, а также некоторые наши работы и работы О. И. Писарик, посвященные развитию положений теории обучающей виртуальной среды (Горожанов, 2024; Gorozhanov, 2019; Pisarik, 2024; Писарик, 2019).

Практическая ценность работы заключается в том, что её результаты могут быть использованы для повышения уровня автоматизации процесса обновления оценочных материалов по иностранным языкам, наполнения лингвистических онлайн-курсов тестовыми и тренировочными заданиями.

Обсуждение и результаты

Подсистема резервного копирования является основополагающей частью всякой системы управления обучением. Рассмотрим здесь и далее пример *LMS Moodle* как «эталонной» системы управления обучением (Горожанов, 2024, с. 42).

Резервное копирование возможно как на уровне всей системы, так и ее отдельных элементов: онлайн-курсов и компонентов онлайн-курсов. Рассмотрим возможности резервного копирования тестовых заданий банка вопросов. Мы выделяем последние среди прочих элементов системы (например, глоссария или базы данных) ввиду их важности. Наш опыт показывает, что составление тестовых заданий и их внесение в систему отнимают значительную долю времени при разработке онлайн-курса.

По этой причине, как правило, авторы ограничиваются минимальным количеством тестовых заданий и не обновляют их в процессе эксплуатации онлайн-курса. Кроме того, тестирование является одной из самых сильных сторон любой системы управления обучением, поэтому технически возможен вариант построения онлайн-курса, состоящего из одних только тестовых модулей для оценивания уровня знаний обучающихся или для их тренировки (режим тренажера) и используемого для организации самостоятельной работы студентов.

В рассматриваемой нами версии *LMS Moodle* 3-го поколения резервное копирование тестовых заданий возможно в следующих форматах:

- формат *Aiken* (подходит только для заданий на множественный выбор и при импортировании выдает много ошибок);
- формат *GIFT* (предусматривает наличие специального метаязыка описания тестовых заданий различных типов и, на наш взгляд, излишне сложен для автоматической генерации; к тому же, для этого формата существует достаточно много генераторов в рамках продуктов *OpenOffice* и *Excel*);
- формат *Moodle XML* (максимально четко структурированный формат, специально разработанный для *LMS Moodle* подвид *XML*; позволяет включать в файл *LMS Moodle* необходимые разработчику теги *HTML*; позволяет резервировать и затем безошибочно разворачивать в системе практически все типы тестовых заданий);
- формат *XHTML* (не является универсальным; фактически только представляет тестовые вопросы в читаемом человеком виде).

Набор форматов импорта несколько отличается:

- вложенные ответы (еще один формат со специализированным метаязыком; поддерживает только тип тестового задания *Cloze Test*);
- формат пропущенного слова (достаточно простой формат для написания кода вручную, подходит только для заданий на пропуски);
- формат *Aiken*;
- формат *Blackboard* (формат для резервных копий *LMS Blackboard*);
- формат *Examview* (формат для резервных копий *Examview 4*);
- формат *GIFT*;
- формат *Moodle XML*;
- формат *WebCT* (формат для резервных копий *LMS WebCT*).

Мы видим, что экспортировать и импортировать тестовые задания возможно только в трех форматах (*Aiken*, *GIFT* и *Moodle XML*), из которых выберем последний по указанным выше причинам – максимальная универсальность и удобство программной обработки.

В качестве примера рассмотрим тестовые задания на заполнение пропусков в тексте из предложенных вариантов. В резервном файле формата *Moodle XML* для этого типа задания можно выделить следующие основные структурные элементы:

1. Тип тестового задания.
2. Название тестового задания.
3. Текст тестового задания.
4. Общий отзыв.
5. Оценка по умолчанию.
6. Штрафные баллы.
7. Флажок перемешивания вариантов ответа.
8. Отзыв при правильном решении.
9. Отзыв при частично правильном решении.
10. Отзыв при неправильном решении.
11. Варианты ответов.

В ходе нашего эксперимента мы заполним блоки 1-3, 5, 7 и 11.

Что касается обучения немецкому языку, который мы задействуем в рамках нашего исследования, то задания на заполнение пропусков могут широко применяться в качестве грамматических (например, выбор правильной грамматической формы), лексических (выбор подходящей лексической единицы из предложенных, выбор синонима/антонима и пр.) и лексико-грамматических (например, выбор подходящего слова в правильной форме).

Определим, возможно ли использовать авторские сбалансированные корпусы для генерации тестовых заданий на пропуски немецкого артикля и окончаний прилагательных.

Алгоритм технических действий для генерации первого вида задания будет следующим:

1. Выбрать диапазон предложений для одного задания (например, 1-10).
2. Выявить по признаку «часть речи» артикли в определенном диапазоне.
3. Сформировать список всех возможных форм артикля для немецкого языка либо тех форм, на которые ориентировано задание.
4. Сравнивая по очереди артикли в тексте (диапазоне предложений), присваивать артиклям номера, соответствующие номеру этой формы в списке +1.
5. Выполнить сборку файла *XML*, добавляя к наименованию файла временную метку.

Для заданий на выбор окончания прилагательных алгоритм будет, по сути, таким же, но вместо списка артиклей необходимо будет сформировать список возможных окончаний прилагательных, а при их нумерации заменять в тексте не прилагательное целиком, а только его окончание.

Поскольку реляционная база данных рассматриваемого корпуса состоит из таблиц предложений и токенов (слов и знаков препинания), то выбрать диапазон предложений не составляет труда. Выборка по частям речи также возможна, поскольку корпус имеет морфологическую разметку, которая, в нашем случае, включает в себя частеречную разметку. Формирование всех возможных форм артикля для того или иного языка не представляет проблемы, так как набор вариаций редко превышает несколько десятков единиц, поэтому

может быть составлен вручную – тем более, что сделать это необходимо лишь единожды. То же самое касается и списка всех возможных окончаний прилагательных, будь то немецкий язык или какой-либо другой, предусматривающий подобный морфологический признак.

Далее, независимо от того, будет это форма артикля, окончание прилагательного, суффикс, приставка или иное, технически необходимо будет выполнить перебор (итерацию) токенов выбранного диапазона стандартным циклом *for* для наполнения формализованной структуры файла *Moodle XML*.

Таким образом, мы приходим к выводу, что наш авторский сбалансированный лингвистический корпус совместим с поставленными задачами.

Следующим шагом исследования явилось написание программного кода на языке *Python*, причем полученный программный продукт (модуль) был интегрирован в упомянутый выше корпусный менеджер и в экспериментальном режиме был помещен в графический интерфейс пользователя в виде пункта меню «Учебное» –> «Комплекс упражнений НЯ» (см. Рисунок 1).

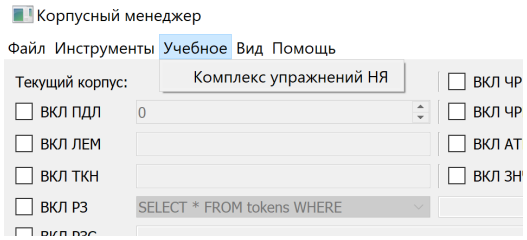

Рисунок 1. *Фрагмент графического интерфейса пользователя: пункт меню для генерации файлов Moodle XML*

В результате на основе первых десяти предложений базы данных корпуса романа «Возлюби ближнего своего» формируется файл *Moodle XML*, который затем импортируется в банк вопросов системы. Просмотр задания показал следующее (см. Рисунок 2).

Рисунок 2. *Представление задания на выбор правильных форм артикля (вариант с правильными ответами)*

Увиденное позволяет заключить, что задание было сформировано безошибочно. Балл по умолчанию был указан равным единице, флажок перемешивания вариантов ответов установлен в позицию «да».

То же было проделано и для задания на окончания прилагательных, но уже на материале романа «На Западном фронте без перемен» (см. Рисунок 3).

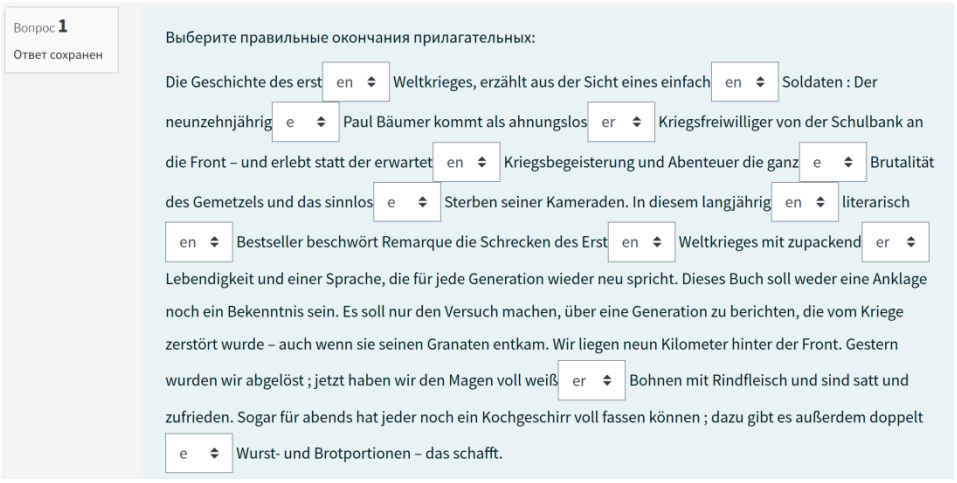

Рисунок 3. *Представление задания на выбор правильных окончаний прилагательных (вариант с правильными ответами)*

Здесь также не было зафиксировано ни технических, ни смысловых ошибок. В обоих случаях написанная программа правильно находила нужные части речи и формировала файл *Moodle XML* полностью в автоматическом режиме.

Несмотря на то, что для эксперимента был выбран диапазон 1-10 предложения корпусов, возможен вариант выбора меньшего/большего диапазона или вариант генерации какого-либо другого количества заданий из всего корпуса или его части, причем так же в полностью автоматическом режиме. При этом количество полученных за несколько минут упражнений может составлять десятки или даже сотни. Для создания значительных объемов типовых упражнений в качестве тренировочных заданий для большого количества обучающихся целесообразно использовать динамические корпусы СМИ, включающие более миллиона токенов (Степанова, 2023).

Поскольку методически правильным является предоставление обучающимся не отдельных упражнений, а комплексов, включающих, например, подготовительный, основной и закрепительный этапы, рациональным кажется снабдить упражнения списком вокабуляра для снятия лексических трудностей до выполнения упражнений.

Возможности корпусного менеджера легко позволяют это сделать, так как включают функцию составления общего частотного списка токенов и частотного списка токенов по частям речи, отсортированных по частоте употребления. На наш взгляд, важнейшими здесь являются существительные, глаголы и прилагательные.

Например, если бы мы составляли такой частотный список для романа «Возлюби ближнего своего», то с помощью запроса к одноименному корпусу получили бы следующее (приводятся первые 20 записей, вспомогательные глаголы и имена собственные исключены):

1) sagen: 914 (говорить);
2) sehen: 567 (видеть);
3) gehen: 472 (идти);
4) kommen: 412 (приходить);
5) fragen: 345 (спрашивать);
6) haben: 305 (иметь);
7) stehen: 278 (стоять);
8) Mann: 270 (человек; мужчина);
9) geben: 248 (давать);
10) Frau: 227 (женщина);
11) machen: 222 (делать);
12) wissen: 191 (знать);
13) nehmen: 184 (брать);
14) Hand: 182 (рука);
15) Tag: 163 (день);
16) bleiben: 162 (оставаться);
17) erwidern: 154 (возражать);
18) Gesicht: 152 (лицо);
19) weiß: 144 (белый);
20) denken: 141 (думать).

При небольшой модификации указанная функция сможет генерировать частотный список на основе заданного диапазона предложений, совпадающего с диапазоном для формирования тестовых заданий, что демонстрирует значительный лингводидактический потенциал нашего корпусного менеджера.

Разработанный в рамках настоящего исследования программный модуль объединяет две функции: для артиклей и для прилагательных – и имеет относительно небольшой объем (ок. 100 строк программного кода в исполняющей части и несколько десятков строк в графической части), что составляет примерно 5% листинга «Генератора сбалансированного лингвистического корпуса и корпусного менеджера». Этот факт также позволяет говорить о возможности относительно быстрого расширения списка функций с целью удовлетворения текущих потребностей образовательного процесса того или иного вуза, а также об удобстве выбранного алгоритма.

Заключение

Итак, мы можем заключить, что поставленные в исследовании задачи были выполнены. Во-первых, был проведен анализ структуры резервных файлов *LMS Moodle*, в ходе которого было установлено, что универсальным и более удобным для программной обработки является формат *Moodle XML*. Во-вторых, была проведена оценка возможностей авторского сбалансированного лингвистического корпуса с морфологической разметкой для автоматической генерации образовательного (лингводидактического) контента для *LMS Moodle*. Корпус показал себя гибким инструментом, поскольку позволяет работать как на уровне токенов, так и на уровне предложений, выделять части речи и составлять частотные списки. В-третьих, нами была написана программа для экспериментальной генерации электронных учебных материалов файлов формата *Moodle XML* – подстановочных упражнений на формы немецкого артикля и окончаний прилагательных. Полученные файлы были импортированы в онлайн-курс и проверены на наличие технических ошибок и правильность учебного содержания.

Заключим, что проведенное прикладное исследование вносит вклад в решение обозначенных нами проблем создания отечественного программного обеспечения и учёта при разработке программных решений принципов методики обучения.

Перспективой исследования является получение методов генерации упражнений на тренировку других языковых явлений, как для немецкого, так и для других иностранных языков, а также для русского языка как иностранного. Далее, необходимо продумать и реализовать генераторы комплексов упражнений, соответствующие требованиям актуальных методов обучения иностранным языкам, которые бы предусматривали не только предварительную работу с вокабуляром, но и задания открытого типа, например, комментарий фрагмента с использованием определенного вокабуляра или развернутые письменные или устные ответы на вопросы к тексту.

Источники | References

1. Горожанов А. И. Подготовка будущих преподавателей иностранного языка к работе в условиях обучающей виртуальной среды (на примере учебного пособия) // Вестник Московского государственного лингвистического университета. Образование и педагогические науки. 2024. Вып. 1 (850).
2. Горожанов А. И., Степанова Д. В. Составление сбалансированного корпуса художественного произведения (на материале романов Ф. Кафки) // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2022. № 7 (862). https://doi.org/10.52070/2542-2197_2022_7_862_31
3. Донина О. В. Технологии искусственного интеллекта в языковом образовании // Журнал филологических исследований. 2023. Т. 8. № 3.
4. Евстигнеев М. Н. Нейросеть Twee – новый инструментарий для педагога английского языка // Вестник Тамбовского университета. Серия «Гуманитарные науки». 2023. Т. 28. № 6. https://doi.org/10.20310/1810-0201-2023-28-6-1428-1442
5. Зубов А. В. Корпусная лингвистика: возможности и перспективы // Русский язык: система и функционирование (к 80-летию профессора П. П. Шубы): материалы III международной научной конференции (г. Минск, 6-7 апреля 2006 г.): в 2-х ч. Мн.: Республиканский институт высшей школы, 2006. Ч. 1.
6. Котюрова И. А. Опыт использования лингвистического корпуса на практическом занятии по немецкому языку // Педагогические мастерские: сборник научных трудов. Киров: Межрегиональный центр инновационных технологий в образовании, 2023. Вып. 23.
7. Лаптев В. В., Ларченкова Л. А., Шубина Н. Л. Системы искусственного интеллекта (ИИ) с Generative Pre-trained Transformer (GPT) архитектурой в формальном, неформальном и информальном образовании // Научное мнение. 2023. № 11.
8. Личаргин Д. В., Бачурина Е. П. Разработка гибкой системы генерации учебных электронных курсов по дисциплине «Иностранный язык» // Информатизация образования и науки. 2021. № 3 (51).
9. Писарик О. И. Модель сопровождающего онлайн-курса английского языка для студентов строительных направлений подготовки // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2019. № 11 (827).
10. Потапова Р. К. Новые информационные технологии и лингвистика: учебное пособие. Изд-е 7-е. М.: URSS, 2021.
11. Степанова Д. В. Программный комплекс для генерации динамического корпуса текстов СМИ // Вестник Минского государственного лингвистического университета. Серия 1 «Филология». 2023. № 6 (127).
12. Сысоев П. В., Клочихин В. В. Формирование коллокационной компетенции студентов на основе корпусных технологий // Перспективы науки и образования. 2022. № 4 (58). https://doi.org/10.32744/pse.2022.4.19
13. Шалаев А. П. Цифровые стандарты – новый этап развития стандартизации? // Стандарты и качество. 2019. № 7.
14. Gorozhanov A. I. Institutional Educational Virtual Environment for Linguistic Purposes: Theory and Practice. Казань: Бук, 2019.
15. Pisarik O. I. Training in English For Construction Applying Web 2.0 Tools as Part of Advanced Course for BIM Designers // Человек – язык – компьютер. Исследователи будущего: материалы научно-практической (заочной) конференции с международным участием (г. Москва, 25 декабря 2023 г.). М.: МГЛУ, 2024.

Информация об авторах | Author information

RU **Горожанов Алексей Иванович**[1], д. филол. н., доц.
[1] Московский государственный лингвистический университет

EN **Gorozhanov Alexey Ivanovich**[1], Dr
[1] Moscow State Linguistic University

[1] *a_gorozhanov@mail.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 07.02.2024; опубликовано online (published online): 25.03.2024.

Ключевые слова (keywords): корпусная лингвистика; лингводидактический потенциал; LMS Moodle; формат Moodle XML; банк вопросов; немецкий язык; corpus linguistics; linguodidactic potential; Moodle XML format; question bank; German language.