САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНСТИТУТ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ РАН
РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ
ИМ. А. И. ГЕРЦЕНА

ТРУДЫ
МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ
«КОРПУСНАЯ ЛИНГВИСТИКА–2019»

24–28 июня 2019 г., Санкт-Петербург

ИЗДАТЕЛЬСТВО САНКТ-ПЕТЕРБУРГСКОГО УНИВЕРСИТЕТА
2019

ББК 81.1
Т78

Ответственный редактор издания
В. П. Захаров

Труды международной конференции «Корпусная лингвистика–
Т78 2019». — СПб.: Изд-во С.-Петерб. ун-та, 2019. — 448 с.

Сборник содержит материалы докладов, представленных на Международной научной конференции «Корпусная лингвистика-2019» 24–28 июня 2019 г. в Санкт-Петербурге.

Создание корпусов текстов является одним из приоритетных направлений в современной лингвистике. Проведение конференции по данной тематике знакомит ученых с современными разработками и новыми технологическими решениями в этой области, а также способствует обобщению опыта научных исследований по корпусной лингвистике.

ББК 81.1

М. Копотев, А. Катинская, С. Иванова, Р. Янгарбер
M. Kopotev, A. Katinskaia, S. Ivanova, R. Yangarber

REVITA: ИЗУЧЕНИЕ ЯЗЫКА НА ОСНОВЕ КОРПУСНЫХ ПОДХОДОВ

REVITA: CORPUS-BASED LANGUAGE TEACHING TOOL

Аннотация. Статья посвящена описанию системы Revita, которая создается в Хельсинском университете. Система представляет собой новаторский подход к проведению индивидуальных тестов и индивидуализированных упражнений, для создания которых активно используются корпуса и инструменты автоматического анализа текста. Данные, собранные в процессе использования системы, открывают путь к индивидуальному подходу в изучении языка, к описанию индивидуальной грамматики ученика.
Ключевые слова. корпусные методы в преподавании языка, компьютерные средства обучения, тестирование.

Abstract. This article describes Revita, a system for assisting language learners, being developed at the University of Helsinki. The system employs a novel approach to progress assessment of learners of foreign languages, and uses both corpus data and NLP tools to automatically generate randomized exercises targeting the learner's level of competency. The data collected from L2 learners offers opportunities and challenges in studying individual grammar from both applied and theoretical perspectives.
Keywords. computer-assisted language learning (CALL), intelligent tutoring systems (ITS), NLP tools, corpus-based

0. Изучение языка с помощью компьютера (англ. computer-aided language learning, CALL) было предложено уже в 50-ые годы прошлого века и тех пор существенно продвинулось благодаря стремительному развитию информационных технологий (Hart 1981; Carol&Jamieson 1983). В настоящее время существуют сотни программ для изучения языка — бесплатных и коммерческих, простых и сложных, для начинающих и экспертов. Некоторые программы не просто используют компьютер как средство обучения, но опираются на современные достижения в области корпусной лингвистики и автоматической обработки языка (Nagata 2002, Heift 2001). Такие системы могут быть названы интеллектуальными обучающими системами. Одной из таких систем посвящена настоящая статья.

1. REVITA (revita.cs.helsinki.fi)

Revita является открытой интернет-платформой, предназначенной прежде всего для поддержки исчезающих языков путем включе-

ния учащегося в активное освоение или развития языковых навыков (Katinskaia et al. 2018). Инструмент предназначен прежде всего для людей, которые уже обладают определенной компетенцией в языке, и не подходит для начинающих изучать язык. Модуль русского языка является самым продвинутым, на его примере мы представим всю систему. Модуль состоит из двух частей, тесно взаимосвязанных и создающих основу для контроля и развития индивидуальной грамматики: система проверки языковой компетенции и языковые упражнения.

1.1. Проверка языковой компетенции

Тесты предназначены для контроля прогресса языковых навыков. Первоначально они были разработаны на Отделении современных языков Хельсинкского университета для использования в процессе обучения финских студентов (Мустайоки 2001, Копотев 2010). В проекте Revita они адаптированы к нуждам студентов с любым родным языком. В своей методической части тесты опираются на процессинговую модель в обучении (англ. processing model) и восходят к так называемым тестам прогресса (англ. progress tests), которые давно применяются, например, в учебных программах подготовки врачей. Основные особенности созданного нами теста, которые выделяют его из многочисленных способов тестирования языковых знаний, описаны ниже.

Задания для студентов выбираются из объемной базы данных, содержащей в настоящий момент около 3,5 тыс. заданий по разным темам: от склонения и спряжения до правописания. Перед тестированием преподаватель и/или студент может определить, какое количество заданий и по какой теме следует выбрать; существует возможность контролировать их общее количество, время, отведенное на ответы, и некоторые другие параметры. Выбор конкретных вопросов из базы осуществляется компьютером автоматически. В настоящий момент тест позволяет проверять знания по следующим разделам:

склонение прилагательных,
склонение существительных,
склонение числительных,
спряжение глаголов,
употребление глагольного вида,
глаголы движения,
глагольное управление,
употребление форм глагольных времен,

устойчивые фразы,
порядок слов,
ударение,
лексическая сочетаемость,
синтаксические конструкции,
орфография.

Каждый раздел, в свою очередь, представлен развернутым списком тем. Все задания в тесте представляют собой вопросы с несколькими вариантами ответов — так называемый *multiple choice* — самая распространенная на сегодняшний день форма, применяемая во множестве тестов (Kehoe 1995). Каждая сессия содержит определенное преподавателем количество заданий, на выполнение каждого отводится по умолчанию 15 секунд (последнее сделано для того, чтобы учащийся не имел возможности проверить ответ).

В ходе тестирования задания даются в случайном порядке, однако в конце теста студент получает распределенные по темам и уровням сложности результаты, которые сохраняются в базе данных, что позволяет студенту уже во время второго тестирования получить сведения о прогрессе своих языковых навыков. Результат содержит как обобщенную оценку — процент правильных ответов, так и детальную картину ответов по всем темам, включенным в сессию.

Безусловно, на выполнение теста влияет и темп выполнения заданий, и внимательность, и усталость от теста. Именно поэтому по каждой микротеме предлагается не менее трех вопросов, что дает возможность получить более надежные результаты. Наконец, важно отметить, что тест можно сдать в любом месте в любое удобное время. Конечно, в таких условиях существует возможность для разного рода манипуляций, но, как показала практика, ограничение по времени и ясная мотивировка приводят к тому, что студенты проходят тесты самостоятельно и без подсказок. Одним из главных преимуществ разработанной системы тестов является возможность продолженной оценки уровня знаний. Такой тонкий контроль динамики позволяет не только оценить прогресс в освоении, но и предложить индивидуальный набор заданий, за которые отвечает второй модуль нашей системы.

1.2. Генерация индивидуальных упражнений

1.2.1. Основная идея этой части системы — стимулирование активного использования языка на основе текстов. Под этим мы подразумеваем активное продуцирование языковых форм в контексте, а не пас-

сивное впитывание языковых примеров или правил. Система предназначена для изучения и грамматики, и лексики, и орфографических правил. Упражнения, которые предлагает система, включают тесты на заполнение пропусков, выбор правильного ответа, кроссворды, задания на карточках и т. д. Ключевой особенностью системы является возможность загрузить тексты по выбору и создать собственную библиотеку. Мы считаем, что индивидуальный выбор текстов значительно повышает мотивацию студента, его вовлеченность в процесс обучения. В то же время студенты и учителя могут делиться текстами или использовать заранее подготовленную библиотеку.

По сути, система дает возможность создать собственный корпус текстов разной степени сложности и автоматически обрабатывает его с помощью стандартных инструментов языкового анализа: морфологического, синтаксического и коллокационного. Для этого используются различные инструменты компьютерной обработки текста, основная обработка русских текстов осуществляется с помощью морфологического анализатора Crosslator (Klyshinsky et al. 2011). На основе выполненного анализа система автоматически создает упражнения: карточки, кроссворд, грамматические упражнения и т. д., которые мы опишем ниже.

1.2.2. Первый тип заданий — **грамматические упражнения**. Для создания упражнений для всех токенов в тексте делается автоматический морфологический анализ и лемматизация. После этого система извлекает из текста токены, которые становятся основой для упражнений. Темы для упражнений могут быть заданы учителем или самим учащимся, однако гораздо более полезной является возможность их генерации на основе тех результатов, которые студент получил в ходе выполнения теста. Таким образом студент получает индивидуальный набор упражнений, которые создаются на основе фрагмента текста объемом в 10–30 слов, или 1–5 предложений. В полученном упражнении часть слов убрана и заменена одним из двух типов заданий: выбор ответа из списка или генерация правильной формы на основе показанной начальной формы.

При создании грамматических заданий существует одно существенное ограничение: токены, которые не могут быть однозначно приписаны одной лемме, на данный момент игнорируются: например, форма *жил* омонимична и имеет два морфологических разбора: прошедшее время глагола *жить* и родительный падеж множественного числа существительного *жи́ла*. Такие случаи игнорируются и не вы-

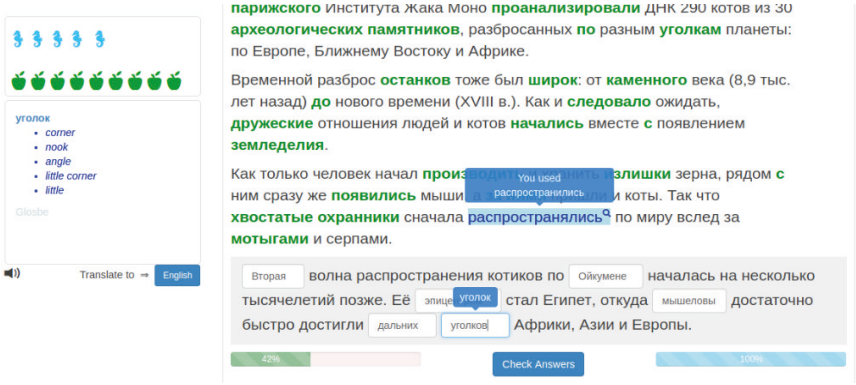

Рис. 1. **Пример упражнения на заполнение пропусков**

бираются для дальнейших упражнений. Система анализирует выполнение упражнений и, основываясь на индивидуальных результатах, предлагает упражнения разной степени сложности, например, выбор правильной формы из предложенных или создание формы при заданной лемме, или создание правильной формы без подсказок. Все ответы, правильные и неправильные, сохраняются в системе. Персональная история учащегося служит основой для выбора заданий для него: например, упражнения, на которые студент никогда не отвечает правильно, отмечаются, но не предлагаются студенту в текущую сессию; упражнения, на которые студент отвечает иногда правильно, иногда нет, получают приоритет и т. д. Для каждого нового задания создаются новые наборы упражнений, так что у студента нет возможности просто вызубрить формы в контексте.

1.2.3. Второй тип — **карточки**, предназначенные для изучения новых слов. Выполняя все виды упражнений, студент может нажать на слово и получить его перевод на выбранный язык. Слова, перевод которых запрашивает ученик, рассматриваются как незнакомые и требующие дополнительного внимания. Все они включаются в набор карточек, связанных текстом, в котором они встретились, а также в набор всех карточек для изучаемого языка. Каждая карточка содержит слово с одной стороны и его сохраненный перевод с другой. Revita позволяет просто листать карточки, а также выполнять упражнения, а именно: ученик должен правильно вставить слово на основе полученного перевода или правильно вставить перевод полученного слова. Пример второго типа упражнений с карточками можно увидеть на Рисунке 2.

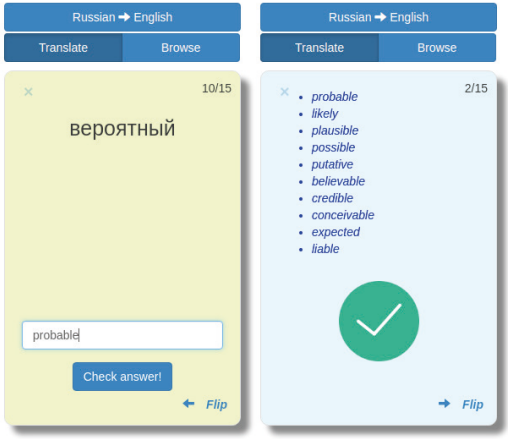

Рис. 2. **Упражнение с карточками**

1.2.4. Третий тип заданий, **кроссворд**, представляет собой лексические упражнения, автоматически создаваемые на основе выбранного текста. Кроссворд состоит из 40–50 слов, выбранных из изучаемого текста. Задача учащегося — вписать в клеточки кроссворда слова в правильной форме. Если учащийся вводит слово правильно, оно добавляется в текст. Если студент испытывает затруднения, система предлагает ему подсказку в виде перевода на известный ему язык. Пример кроссворда можно увидеть на Рис. 3.

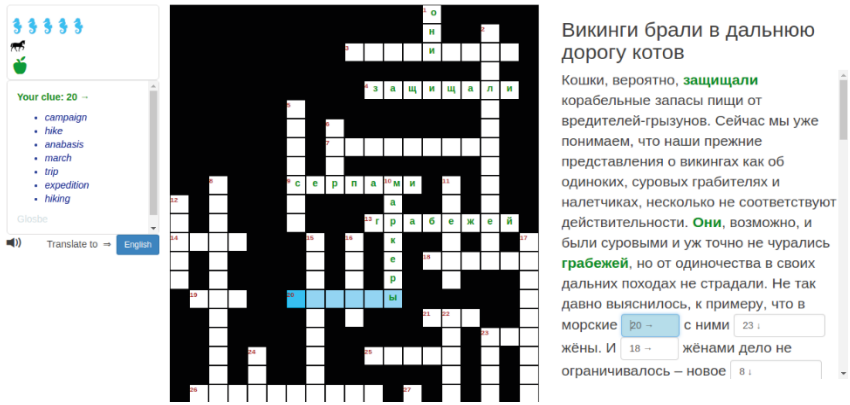

Рис. 3. **Кроссворд**

На любом этапе, обычно после изучения темы, учащийся снова проходит тест (целиком или только ту тему, которую он изучил) и получает сведения о своем прогрессе в освоении языка. Результаты представляются в виде диаграмм на специальной странице: чем больше круг на диаграмме, те больше заданий выполнил студент. Чем больше правильных ответов дал студент, тем более ярким становится этот круг.

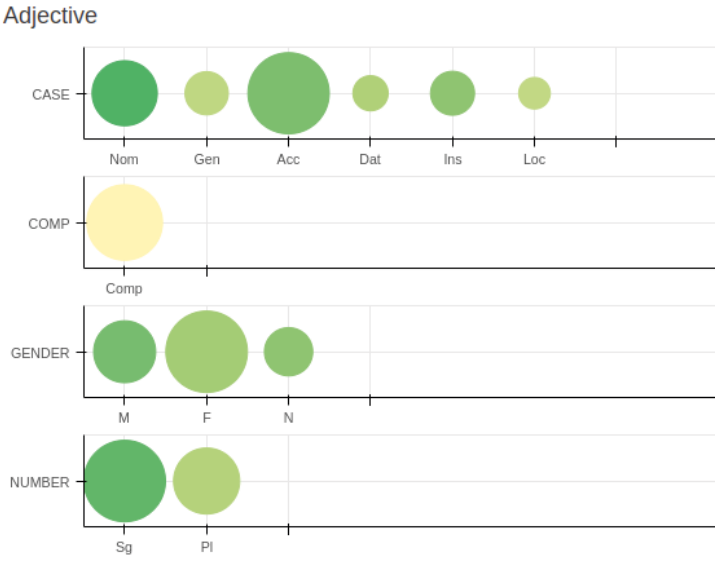

Рис. 4. **Прогресс языковых навыков**

На Рисунке 5 показана общая схема тонкой настройки заданий, исходя из действий конкретного пользователя. После прохождения теста (assessment module) строится модель индивидуальных навыков (student module) с учетом конкретных языковых тем (domain module). Это становится основой для индивидуализированных заданий (exercise module), результаты которых хранятся в истории (student history). На каждом следующем шаге система подстраивается под нужды студента, учитывая его прогресс в выполнении конкретных заданий. Система в принципе работает автономно, однако у преподавателя есть возможность настраивать ее в ручном режиме (instruction module), задавая темы, тексты, параметры тестирования и т. д.

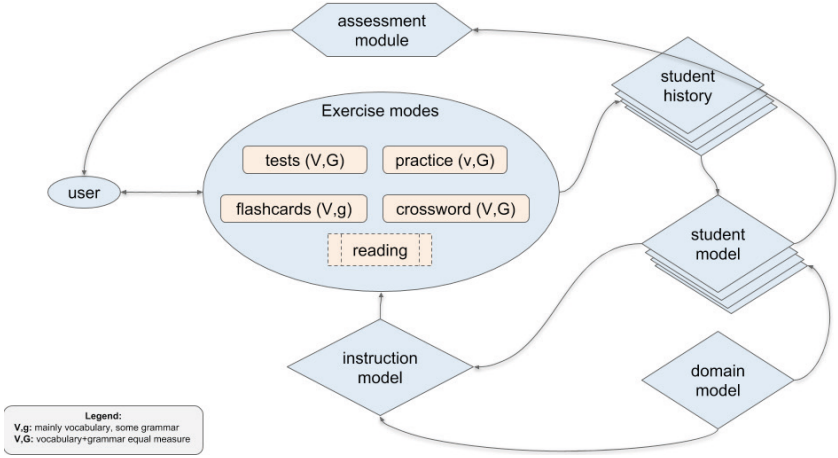

Рис. 5. **Общая схема работы системы**

2. Выводы

На сегодняшний день в лингвистике накоплен солидный опыт по обработке текстов и созданию разнообразных корпусов. Существуют коллекции текстов и инструменты автоматического анализа для множества языков. Следующий шаг состоит в том, чтобы создавать ресурсы, опирающиеся на эти достижения, один из которых представлен в настоящей статье. Система Revita находится в стадии разработки. Однако уже сейчас ясно, что использование достижений компьютерной лингвистики открывает потенциал для изучения и преподавания языка в индивидуальном режиме. Изучающему Revita дает возможность создавать бесконечное количество заданий с учетом собственных потребностей, преподавателю — возможность гибкого динамичного контроля результатов как одного студента, так и всей группы, объединенной общей грамматической темой. Кроме того, постепенное накопление знаний о том, какие темы являются сложными или легкими для всех изучающих язык открывают путь к созданию учебных материалов, исходящих из реальных проблем, с которыми сталкивается студент-иностранец или носитель эритажного языка.

Список литературы

1. *Chapelle, C. and Jamieson, J.* (1983), Language lessons on the PLATO IV system, System, 11(1), pp. 13–20.
2. *Klyshinsky, E. S., Kochetkova, N. A., Litvinov, M. I., Maximov, V. Yu.* (2011), Method of POS disambiguation using information about words cooccurrence (for Russian), Proc. of GSCL, pp. 191–195.
3. *Hart, R.* (1981), Language study and the PLATO system, Studies in Language Learning, 3(1), pp. 1–24.
4. *Heift, T.* (2001), Intelligent language tutoring systems for grammar practice. Zeitschrift fur Interkulturellen Fremdsprachenunterricht, 6(2).
5. *Katinskaia, A, Nouri, J., Yangarber, R.* (2018), Revita: a language learning platform at the intersection of ITS and CALL. 11th edition of the Language Resources and Evaluation Conference, 7–12 May 2018, Miyazaki (Japan).
6. *Kehoe J.* (1995), Writing multiple-choice test items, Practical Assessment, Research & Evaluation. Vol. 4 (9).
7. *Nagata, N.* (2002). Banzai: An application of natural language processing to web-based language learning, CALICO journal, pp. 583–599.
8. *Копотев М. В.* (2010), Система прогрессивного тестирования KARTTU (описание и первые результаты), Русский язык за рубежом, 3, с. 23–29.
9. *Мустайоки А.* (2001), Упражнения по русской грамматике: таксономия и база данных, Русский язык на рубеже тысячелетий, СПб, с. 122–138.

References

1. *Chapelle, C. and Jamieson, J.* (1983), Language lessons on the PLATO IV system, System, 11(1), pp. 13–20.
2. *Hart, R.* (1981), Language study and the PLATO system, Studies in Language Learning, 3(1), pp. 1–24.
3. *Heift, T.* (2001), Intelligent language tutoring systems for grammar practice. Zeitschrift fur Interkulturellen Fremdsprachenunterricht, 6(2).
4. *Katinskaia, A, Nouri, J., Yangarber, R.* (2018), Revita: a language learning platform at the intersection of ITS and CALL. 11th edition of the Language Resources and Evaluation Conference, 7–12 May 2018, Miyazaki (Japan).
5. *Kehoe J.* (1995), Writing multiple-choice test items, Practical Assessment, Research & Evaluation. Vol. 4 (9).
6. *Klyshinsky, E. S., Kochetkova, N. A., Litvinov, M. I., Maximov, V. Yu.* (2011), Method of POS disambiguation using information about words cooccurrence (for Russian), Proc. of GSCL, pp. 91–195.
7. *Kopotev M. V.* Sistema progressivnogo testirovaniia KARTTU (opisanie i pervyje rezul'taty) [The progress test KARTTU: description and first results] // Russkij jazyk za rubezhom. 2010. 3, pp. 23–29.
8. *Mustajoki A.* Uprazhnenija po russkoj grammatike: taksonomija i baza dannyx [Russian grammar exercises: taxonomy and database] // Russkij jazyk na rubezhe tysjacheletij. Saint Peterspurg, 2001. 122–138

9. *Nagata, N.* (2002). Banzai: An application of natural language processing to web-based language learning, CALICO journal, pp. 583–599.

Копотев Михаил
Хельсинкский университет (Финляндия)
Mikhail Kopotev
University of Helsinki (Finland)
mihail.kopotev@helsinki.fi

Анисия Катинская,
Хельсинкский университет (Финляндия)
Anisia Katinskaia
University of Helsinki (Finland)
anisia.katinskaia@cs.helsinki.fi

Сардана Иванова
Хельсинкский университет (Финляндия)
Sardana Ivanova
University of Helsinki (Finland)
sardana.ivanova@helsinki.fi

Роман Янгарбер
Хельсинкский университет (Финляндия)
Roman Yangarber
University of Helsinki (Finland)
roman.yangarber@cs.helsinki.fi