Вера Григорьевна СИБИРЦЕВА
Николай Вячеславович КАРПОВ
(Нижний Новгород)

Автоматическая адаптация текстов для электронных учебников. Проблемы и перспективы (на примере русского языка)

Automatic adaptation of the texts for electronic textbooks. Problems and perspectives (on an example of Russian)

The paper is intended to describe the experience of using the authentic linguistic corpus materials within the project "Creating an electronic textbook of Russian as a foreign language". Special attention is paid to the fundamental principles of the new project – automatic adaptation of RNC's linguistic material. Worked out by means of information technologies, the product is supposed to adapt the complexity of authentic texts in terms of their syntactic and morphologic structures and vocabulary. The stages indispensible to attain the objective are also explained in the article. The paper describes not only the algorithm for solving the tasks and the final result of the research, but also the difficulties, which the developers face, and their solutions.

Key words: Automatic adaptation of texts; Russian National Corpus; electronic textbook; Russian as a foreign language; syntactic and morphological structure

В период быстрого развития технологий и международной глобализации русский язык приобретает все большую популярность во всем мире. Для обучения русскому языку создаются учебники как в печатном, так и в электронном формате, хотя до настоящего времени бумажные обучающие материалы значительно преобладали и ощущался недостаток мультимедийных учебников и сборников упражнений. Качество электронных книг во многом зависит от используемых технологий, поскольку сложность электронного учебника не только препятствует усвоению учебного материала, но и вызывает определенное неприятие студента. Еще больший дефицит наблюдается в сфере создания книг для чтения по культуре современной жизни России в целом. Тексты, создаваемые для обучения русскому языку, должны не только соответствовать уровню иностранного читателя, но и быть актуальными, а значит, постоянно обновляться. При создании учебников и книг для чтения на бумажной основе их отставание от современных реалий составляет в среднем 5–10 лет, эту инертность можно было бы преодолеть, создавая электронные базы адаптированных текстов.

Отсутствие достаточного количества мультимедийных учебников и книг для чтения в области русского языка как иностранного при увеличении спроса на них стало отправной точкой для работы научно-учебной группы студентов и преподавателей НИУ ВШЭ «Корплинги».

На первом этапе на материале НКРЯ было создано электронное учебное пособие «Русский глагол» в системе дистанционного обучения e-front. Пособие было сосредоточено на продуктивном префиксальном словообразовании русского глагола и предназначалось для студентов продвинутого уровня обучения (B2, второго сертификационного). Иностранным студентам предлагалось выбрать любой интересующий их префикс из предложенного списка и ознакомиться подробнее с его значениями. Теоретический материал сопровождался примерами из текстов на различные темы. В отличие от подавляющего большинства электронных учебников, данное учебное пособие было полностью разработано на материале современного русского языка. Кроме того, расстановка ударений во всем учебнике позволяет тренировать правильное интонирование предложений и произношение.

Предпочтение Национальному корпусу русского языка при создании учебника было отдано именно потому, что Корпус содержит не только актуальные, но и очень разнообразные языковые материалы. Более чем 500 миллионов словоупотреблений помогает обеспечить разнообразие учебных материалов, связанных не только с художественной литературой, но и научными, деловыми, публицистическими текстами и разговорным языком.

Неоспоримым преимуществом использования НКРЯ для создания новых интерактивных учебников являлась возможность быстро определить наиболее частые грамматические и лексические конструкции. В отличие от формальных, искусственных примеров, которые в значительной степени присущи бумажным учебникам, электронный учебник группы «Корплинги» помогал иностранцам ориентироваться на живой язык и знакомил их с изменениями в языке. Широкий охват лексики, связанной с публицистикой, онлайн-блогами, форумами и прочими сферами коммуникации, позволял обучающимся расширить словарный запас в своей профессиональной сфере.

Данный этап работы показал, что применение аутентичных материалов НКРЯ в качестве базы для учебных пособий очень перспективно, но в то же время содержит немало трудностей. Стремление наполнить новое электронное пособие актуальным материалом и тем самым отразить в учебнике пласт действительно живого современного, а не книжного языка не всегда соответствует учебным задачам. Тексты, которые отражают живой русский язык и преобладают в Корпусе, большей частью взяты из периодики, блогов и форумов. В первую очередь трудности связаны со сложностью лексики и синтаксических конструкций, характерных для неадаптированного русского языка. Электронное учебное пособие «Русский глагол» было ориентировано на продвинутый уровень обучающихся, в нем затрагивались вопросы словообразовательной семантики. Именно поэтому было решено обратиться к неадаптированным текстам, создающимся в наше время.

Задания учебного пособия «Русский глагол» тестировались участниками летней школы по русскому языку как иностранному, проходившей в Нижнем Новгороде в июле 2012 года, а также студентами, обучавшимися в НИУ ВШЭ по обмену в 2012–2013 гг. Пособие

включает в себя объяснительный теоретический материал, упражнения различных типов (в том числе и видеоупражения), а также тестовые задания. Пособие до настоящего времени бесплатно доступно для пользователей в системе дистанционного обучения e-front, которая сочетает в себе функции систем управления обучением и систем управления и создания учебных материалов. Контроль за выполнением заданий также осуществлялся дистанционно. Процесс отбора примеров для теоретической части и упражнений занял значительное количество времени, т. к. разработчики сознательно отказались от адаптации материала, и приходилось просматривать большое количество примеров, чтобы найти соответствующие продвинутому уровню обучения (B2). К сожалению, сложность аутентичной лексики (профессиональные, многозначные слова) все же была выше уровня обучающихся, которые приезжали в основном со знаниями порогового уровня (B1).

Необходимость адаптировать современные тексты в обучающих целях и желание автоматизировать процесс адаптации легли в основу второй части проекта, связанной с созданием базы адаптированных публицистических текстов. Методы сбора, обработки и классификации языковых примеров, а также разработка компьютерной оболочки для представления продукта в интерактивном, дружественном к пользователю виде, освоенные в ходе работы над проектом «Русский глагол», стали закономерным этапом перехода от научно-прикладных задач к экспериментально-научным.

В настоящее время пользователями в интернете генерируется большое количество контента, на любом языке мира. Среди этого контента новости, аналитические новостные статьи, комментарии пользователей и многое другое. Использование текстов подобного типа для обучения иностранному языку представляется интересным и перспективным, так как чтение на иностранном языке формирует навык восприятия новой и актуальной информации на другом языке. Понимание содержания статей развивает обучающихся, продуктивность данной деятельности неоспорима не только в образовательных, но и в познавательных целях. Доступность аутентичного материала в интернете приводит к тому, что уже начиная с порогового уровня владения

языком (B1), многие обучающиеся стремятся читать не только тексты учебников и книг для чтения.

Основным препятствием использования актуальных публицистических текстов не носителями языка в обучении является высокая сложность таких текстов. Авторы в большинстве своем ориентируются на читателя, для которого язык является родным, а при обучении чтению на иностранном языке важно, чтобы уровень текста соответствовал уровню подготовки обучающегося. Чтобы привести текст в соответствие с уровнем владения языком, нужно значительным образом переработать текст, что порой не проще, чем написать его заново. Такой процесс называется адаптацией и производится вручную. Адаптированные тексты, несомненно, легче воспринимаются читателем, не носителем языка.

Скорость адаптации текста даже специалистом в области обучения языкам достаточно низкая. Это обуславливает высокую стоимость такого процесса и большие временные затраты, а потому неприменимо к большому объему текстов, существующих в сети. Поэтому автоматизация процесса адаптации текста представляется интересной задачей.

Исследование именно новостных текстов является перспективным по еще одной причине. Поисковые системы учитывают частоту обновления и появление новой информации для ранжирования интернет-ресурсов. Этот факт и желание находиться на верхних позициях в ранге поисковиков, породили целую индустрию, которую называют рерайтинг. Задача человека, занимающегося рерайтингом, – написать статью на тему последней новости. При этом он узнает новость из сообщений информационных агентств и других открытых источников. Рерайтер оперирует теми же фактами, пишет про тех же персон, географические объекты и прочее, но подбирает выражения и выстраивает текст таким образом, чтобы он не дублировал источник информации. В результате в сети генерируется большое количество текстов, в основе которых лежит схожий информационный повод. Использование этого факта для подбора синонимичных слов и выражений из новостей выглядит многообещающе.

Целью работы «Корплингов» на втором этапе являлось исследование подходов для автоматизации процесса адаптации текста в новост-

ном жанре. С промежуточными результатами работы можно ознакомиться на тестовой странице: Пользователь может ввести неадаптированный текст и лексика, выходящая за пределы выбранного уровня, будет подсвечена красным цветом. Программа для выявления уровня лексики написана студентами, участниками группы, на языке Python. Поиски путей адаптации текстов начались с глубокого исследования синтаксической структуры и лексической сложности предложений в существующих учебниках по русскому языку как иностранному. Результаты показали, что эти предложения адаптированы для обучающихся с очень ограниченным знанием языковых структур. Например, большинство текстов до уровня B1 избегают причастных и деепричастных оборотов, не содержат информацию в скобках, не показывают всех формальных возможностей выражения прямой речи и т. д. Словари-глоссарии включают в себя только активный лексический минимум, отвечающий определенному уровню компетентности в языке без учета сложных слов. В то же время новостные тексты и примеры НКРЯ представляют собой реальную базу используемого языка, уровень их синтаксической сложности зачастую гораздо выше, чем в существующих учебниках.

Для написания правил упрощения синтаксических структур была проанализирована часть морфологически размеченного Национального корпуса русского языка (Синтагрус). Все предложения в Синтагрусе были разбиты на «простые», соответствующие пороговому уровню владения русским как иностранным, и «сложные» соответственно. Из поставленных на втором этапе задач одна уже достигнута: создана программа, автоматически идентифицирующая лексический уровень сложности введенных данных и маркирующая цветом известную и неизвестную лексику. Вторая задача более сложная: автоматическое преобразование комплекса примеров в «простые» для того, чтобы создать базу адаптированных текстов для электронного учебника. Конечный продукт будет содержать автоматически упрощенные тексты и станет посредником между обучающимися и аутентичными текстами. Выявление правил преобразования и адаптации синтаксической структуры без потери смысла может также дать импульс теоре-

тическим исследованиям в когнитивной области и в области прикладной работы с языком.

Дальнейшие шаги включали в себя подробное изучение механизмов адаптации текста с точки зрения лингвистических правил, работу с лексическими анализаторами, частотными словарями. Запущен процесс программирования дополнений, которые помогают адаптировать и упростить языковые структуры, а также настроить лексические анализаторы.

Для использования текстов в учебных целях обязательна их лексическая, морфологическая и синтаксическая адаптация. С точки зрения лексики, аутентичные тексты могут включать в себя довольно редкие, сложные слова, а также неологизмы и специальную лексику различных сфер деятельности. Было необходимо сопоставить лексическую сложность примеров в традиционных учебниках и новостных примерах: если в первых количество незнакомых лексем не превышает 7–15 % для того или иного уровня владения языком, то последние содержат до 40 % атипичной лексики. Как правило, если объем незнакомых пользователю слов превышает критический уровень, это существенно затрудняет понимание текста и выполнение учебных заданий (к примеру, по грамматике).

Адаптация лексики должна быть произведена следующими методами:

1. Синонимическая замена слов (свыше → более, глава → руководитель)
2. Замена гиперонима на гипоним, родового слова на более частное слово (табачные изделия → сигареты)
3. Замена гипонима на гипероним, выражающий генерализацию или обобщение (врач-терапевт → врач; Путин поймал большую щуку → Путин поймал большую рыбу).
4. Замена анафор в тексте, когда одно и то же лицо называется в тексте и «врач», и «собеседница агентства». Естественно, «врач» в контексте адаптирования «проще», чем «собеседница агентства». Но машине очень сложно определить, что «собеседница агентства» в тексте – это тот же субъект, что и врач X. Реализация такой

замены автоматически потребует мощный модуль разрешения анафор.

5. Удаление суффиксов субъективной оценки (магазинчик → магазин; машинка → машина или автомобиль)

Полностью автоматизированная система процесса адаптации текста должна принимать исходный сложный текст, преобразовывать его по своим алгоритмам, и выдавать упрощенный текст, содержащий информацию с минимальным искажением относительно исходного. Пороговый уровень владения языком был выбран целевым уровнем упрощения, при этом смысловое искажение минимизировалось. В настоящее время осуществлено маркирование лексики и вводных конструкций.

В ряде случаев применение только лексических методов адаптации позволяет значительно упростить восприятие текста. К тому же они достаточно хорошо подходят для автоматической реализации. В соответствии с составленными правилами выбор между наилучшим словом-заменителем и исходным происходит только на основании частотности употребления в языке.

Употребление слова-заменителя иногда требует согласования слов контекста, то есть изменения словоформы согласующихся слов. Например, слово «машина» женского рода, а слово «автомобиль» мужского рода. Производится также морфологическая обработка контекста.

Метод исследования контекста возможен с использованием латентного размещения Дирихле. Это порождающая модель, позволяющая объяснять результаты наблюдений с помощью неявных групп, что позволяет получить объяснение, почему некоторые части данных схожи. Например, если наблюдениями являются слова, собранные в документы, утверждается, что каждый документ представляет собой смесь небольшого количества тем и что появление каждого слова связано с одной из тем документа. Когда человек пишет текст, он отталкивается от некоторого набора тем. Для этого человек сначала выбирает тему, на которую он хочет писать. В рамках этой темы есть распределение слов, которые могут фигурировать в любом документе, где содержится эта тема. Соответственно, в рамках распределения генери-

руется слово документа. Далее ситуация повторяется. Выбирается новая тема либо сохраняется старая и снова генерируется слово в рамках этой темы. Целью модели является построение статистической модели пользователя, который пишет текст и, соответственно, ограничивает степень синонимического расширения контекста, используемого при упрощении.

Для построения автоматизированной системы упрощения эмпирически исследовалось, как человек адаптирует тексты. Был сформирован набор актуальных новостных текстов, различных по своим тематикам. Набор включал в себя 10 текстов информационного агентства «РИА». Затем была проведена работа по ручной адаптации данных публицистических текстов, которая осуществлялась двумя независимыми экспертами, специалистами в области преподавания русского языка как иностранного. Тексты были упрощены до порогового уровня владения русским языком как иностранным. При этом методы, используемые экспертами, протоколировались. По завершению работы был составлен отчет, где все методы были систематизированы.

Неадаптированный текст:

Лишь около трети населения Земли на сегодняшний день имеет возможность доступа в интернет, однако к 2020 году выходить в сеть сможет практически все население планеты – в основном, за счет мобильных устройств, пишет издание Digit.ru со ссылкой на заявление вице-президента по развивающимся рынкам Google Мохаммада Гавдата в ходе Петербургского международного экономического форума. По словам Гавдата, рынок устройств с выходом в интернет кардинально меняется под воздействием нескольких сил – новых типов пользователей, форматов потребления контента и стоимости продуктов и услуг. Примерами тому он называет использование мобильных устройств как «вторых экранов» при просмотре телевизора и получение прямого ответа на поисковый запрос, введенный голосом со смартфона или планшета.

Адаптированный вариант 1:

Только около 30 % людей на сегодняшний день имеет доступ в интернет. Но к 2020 году выходить в интернет сможет почти все население планеты – при помощи мобильных устройств. Рынок устройств с выходом в интернет меняется из-за новых типов пользователей, стоимости про-

дуктов и услуг. Создаются мобильные устройства как «вторые экраны», чтобы смотреть телевизор, и устройства, чтобы получать прямой ответ со смартфона или планшета при голосовом поиске.

Адаптированный вариант 2:

Только около трети населения Земли сегодня имеет доступ в интернет. Однако к 2020 году почти все население планеты сможет пользоваться интернетом. Большинство людей будут использовать для этого мобильные устройства. Об этом сказал вице-президент Google Мохаммад Гавдат на Петербургском международном экономическом форуме. Гавдат сказал, что рынок устройств с выходом в интернет сильно меняется. Появляются новые типы пользователей. Например, некоторые люди используют мобильные устройства как «вторые экраны», когда смотрят телевизор. Изменяется стоимость продуктов и услуг.

После сопоставления полученных результатов был предложен список правил, описывающих способы морфологической адаптации предложений.

Примеры адаптации существительных:

а) существительное/глагол (которые можно заменить на предикатив) и существительное, образованное путем номинализации глагола, заменяются на предикатив + глагол:

требует *долгого* ***тестирования и отладки*** ⇒ ***нужно*** *много* ***тестировать*** *и* ***отлаживать***

б) аббревиатуры и сокращения заменяются полными формами слов или обобщенными синонимами

Минобрнауки ⇒ ***Министерство образования и науки***

соцсети ⇒ ***социальные сети***

ОАО *«Ростелеком»* ⇒ ***компания*** *«Ростелеком»*

в) существительное, образованное путем субстантивации причастия, подлежит замене на конструкцию «тот, кто + глагол»

желающим бросить курить ⇒ **тем, кто хочет** бросить курить

Подобные правила были созданы для всех частей речи.

Разумеется, не представляется возможным сразу достичь необходимых результатов с использованием автоматической адаптации. Учитывая, что работа по автоматизации работы лингвистических правил и их дальнейшей верификации может занимать довольно продолжи-

тельное время, в качестве альтернативного пути был предложен вариант маркирования морфологических единиц, усложняющих синтаксическую структуру предложения. Для исходного текста маркирование осуществляется последовательно в разных режимах, на каждом этапе определенным цветом выделяются «сложные» единицы. Визуальное выделение облегчает ручную адаптацию, которая пока остается самым плодотворным вариантом упрощения текста:

1. Цветом выделяются цепочки существительных в родительном падеже
   о необходимости принятия мер административного характера.
2. Маркируются конструкции, состоящие из глагола в изъявительном наклонении и инфинитива, если между ними нет знаков препинания. Их часто можно заменить одним глаголом.
   позволяет просчитать, может привлечь
3. Выделяются причастные обороты, которые впоследствии можно заменить придаточными предложениями, более легкими для восприятия
   содержащих информацию – которые содержат информацию
4. Разнообразные составные союзы могут быть успешно заменены на более употребительные простые, список их конечен
   не только – но и; как – так и; до тех пор, пока; несмотря на то, что

и пр.

Параллельно велась работа по упрощению синтаксических структур русского языка. Потребовался анализ грамматической системы русского языка в соответствии с нормативными справочниками и пособиями, а также курсами лекций по морфологии и синтаксису русского языка. Путем сравнения материала, который в соответствии со стандартами для изучения русского языка как иностранного на базовом и первом уровнях должен присутствовать в грамматическом минимуме, на базе обобщенной грамматической системы русского языка вручную были выделены структуры (синтаксические, семантические осложнители, а также сложные предложения), которые слишком трудны для восприятия на базовом и первом уровнях владения русским языком.

Важно, что осложняющий компонент может быть выражен со структурной точки зрения любой языковой единицей: отдельной словоформой (чаще всего акцентируется с помощью частиц, союзов и др.); сочинительным рядом; словосочетанием; оборотом (грамматической конструкции из главного слова и зависимых слов); предложением.

Сложные предложения были классифицированы с точки зрения грамматических связей и формальных показателей: наличие нескольких глаголов одной грамматической формы (грамматический показатель), характерные для разных видов связей союзы (формальный показатель) и пр.

После создания классификации была проделана работа по заполнению слотов классификации, то есть непосредственно работа по написанию правил, позволяющих отсекать сложные, не входящие в грамматический минимум конструкции.

Одной из самых важных задач являлось описание коллекции правил, входящих в синтаксический минимум по РКИ первого сертификационного уровня, а также непрерывное пополнение коллекции так называемых «запрещающих» правил (описывающих структуры, которых не должно быть в базе адаптированных предложений). Необходима была их интеграция с правилами, описывающими простые структуры, входящие в синтаксический минимум, в разрезе их наложения друг на друга и улучшения работы программы.

Пример запрещающих правил:

Не должно быть конструкций с формулой:

первое предложение содержит больше, чем 8 словоформ, а затем следует сложносочиненное или сложноподчиненное предложение

более восьми словоформ в предложении, затем союз [и а, но, или] и вторая часть сложносочиненного предложения или сложноподчиненного предложения из более чем 8 словоформ.

Предложений, содержащих экспликаторы субъективной модальности [вообразите, вообразите себе, вообще, вообще говоря, вообще-то]. *Приводится список исключаемых единиц.*

В задачу лингвистов входила необходимость создания коллекции правил, которые в дальнейшем могли быть запрограммированы и использованы в качестве основных:

а) для извлечения из текстового материала различной сложности наиболее простых предложений путем отсечения сложных для восприятия, не входящих в грамматический минимум конструкций;

б) для упрощения сложных для восприятия, не входящих в грамматический минимум конструкций.

После изучения актуальных грамматических справочников и грамматических минимумов для обучения русскому языку как иностранному в соответствии с требованиями к обучению РКИ были выделены типы условно простых предложений (от одночленных до пятичленных). Для каждого типа двучленных и трехчленных предложений были расписаны возможные комбинации морфологической сочетаемости членов предложения. Приведем примеры именных групп:

1. NUMR masc/femn/neut, sing/plur + NOUN anim/inan, masc/femn/ /neut, sing/plur, nomn *Десятый билет*.
2. PRTF masc/femn/neut, sing/plur, nomn + NOUN anim/inan, masc/ /femn/neut, sing/plur, nomn *Стареющая кокетка*.
3. Supr masc/femn/neut, sing/plur, nomn + NOUN anim/inan, masc/femn/ /neut, sing/plur, nomn *Прекраснейшая девушка*.

При разработке лингвистических правил по упрощению предложений (для пополнения материалов электронного учебного пособия) коллектив научно-учебной группы попытался определить способ, по которому будет строиться автоматический алгоритм: нужно было в том числе выбрать, будет ли это путь отсечения сложных для восприятия, не входящих в грамматический минимум конструкций, либо это будет алгоритм, основанный на грамматическом подобии (при наличии образцовых, достаточно простых грамматических структур по их образу и подобию автоматически отбираются схожие). Было принято решение интегрировать два возможных пути, чтобы добиться наилучших результатов.

Построение полуавтоматической системы для адаптации аутентичных текстов в учебных целях – практически ориентированная и одновременно исследовательская деятельность, имеющая широкие пер-

спективы, не только в области обучения, но и в коммерческой деятельности (например, рерайтинга). Необходимо еще раз подчеркнуть, что создание такой системы стало возможным с применением материалов Национального корпуса русского языка. В настоящее время исследователи обычно прибегают к Корпусу в иллюстративных целях, несмотря на то, что НКРЯ отличается удобной лингвистической разметкой, лексическим и грамматическим разнообразием. Разработка предварительного инструментария для адаптации примеров НКРЯ и оригинальных публицистических текстов может быть ключевым шагом, необходимых для расширения и ускорения массового внедрения материалов Корпуса в число образовательных ресурсов, как интерактивных, так и традиционных. Особенно важно подчеркнуть, что описанные методы использования ресурсов НКРЯ не замыкаются только на русском языке и могут с успехом применяться на базе лингвистических корпусов других языков. А это, в свою очередь означает повышение значимости актуального содержания электронных учебников и обеспечение качественного образования в целом.

В данной научной работе использованы результаты, полученные в ходе выполнения проекта «Адаптация языкового материала НКРЯ для электронного учебника „Русский язык как иностранный“», выполненного в рамках Программы «Научный фонд НИУ ВШЭ» в 2013 году, грант № 13-05-0031.

Библиография:

АКИШИНА, А. А., КАГАН, О. Е.: *Учимся учить. Для преподавателя русского языка как иностранного.* 2-е изд., испр. и доп. Москва: Рус. яз. Курсы, 2002. 256 с.

Государственный стандарт по русскому языку как иностранному: Базовый уровень. Москва–Санкт-Петербург: «Златоуст», 2001. 112 с.

НКРЯ: *Что такое Корпус?* http://www.ruscorpora.ru/corpora-intro.html.

ТРЕБОВАНИЯ ПО РУССКОМУ ЯЗЫКУ КАК ИНОСТРАННОМУ: *Первый уровень. Общее владени*е. Москва–Санкт-Петербург: «Златоуст», 2007. 89 с.

DAVID M. BLEI, ANDREW Y. NG, MICHAEL I. JORDAN: *Latent Dirichlet allocation*. Journal of Machine Learning Research. 01 2003, pp. 993–1022.

SIBIRTSEVA, V., KARPOV, N.: Development of modern electronic textbook of Russian as a foreign language: content and technology / Working papers by Издательский дом НИУ ВШЭ. Series WP "Working Papers of Humanities". 2012. No. 2012-6.

Приложения:

АДАПТАЦИЯ ТЕКСТА: http://lingvocourse.ru/www/public_html/cgi-bin/simp/textarea.py

РУССКИЙ ГЛАГОЛ: http://lingvocourse.ru/www/index.php?ctg=lesson_info&courses_ID=1