Л. Н. Беляева

ИНФОРМАЦИОННОЕ ПРОСТРАНСТВО ФИЛОЛОГА И ПРИНЦИПЫ ЕГО ОРГАНИЗАЦИИ

(Статья подготовлена при поддержке фонда РГНФ.
грант 110400/130151 МО-2/03)

Современный филолог должен не только обладать знаниями и умениями в области информационных технологий, но и иметь возможность доступа к специализированным средствам поддержки его учебной, научной и методической работы. Такие средства должны быть направлены на решение конкретных исследовательских и учебных задач (анализ и перевод текстов, исследование текстов, записанных в электронном формате, выполнение учебных заданий) и на проведение научных исследований на базе больших массивов текстов. Обеспечить такую информационную поддержку можно

только путем создания специализированного автоматизированного рабочего места (АРМ) филолога. Подобное АРМ, разработка которого является ясно осознанной необходимостью, представляет собой комплекс программных, лингвистических и технических средств, обеспечивающий удобство работы и потребности как студента, обучающегося по направлениям филологического цикла, так и преподавателей и исследователей в различных отраслях этой области знаний (лингвистов, литературоведов, методистов и т. д.).

Решающим аспектом развития общества в современном мире является научное и культурное взаимодействие, важным средством осуществления которого является компьютерная техника с ее новыми информационными технологиями. Учителя и специалисты в гуманитарных областях знаний, и в первую очередь лингвисты, изучающие естественный язык, который остается основным средством формирования, хранения и передачи информации, оказываются в этой ситуации на переднем крае такого взаимодействия, важность которого сегодня трудно переоценить. При этом особое значение приобретает как языковая и общекультурная квалификация тех, кто призван осуществлять это взаимодействие, так и вопрос о том, насколько специалист-преподаватель готов работать в поликультурном и многоязычном информационном пространстве.

Однако большинство преподавателей-практиков, даже имеющих доступ к информации в системе Интернет, не владеет иностранными языками, и в первую очередь английским языком, в той степени, чтобы свободно воспринимать постоянно увеличивающийся объем информации по новым и развиваемым методам и подходам к обучению. В то же время вхождение нашей страны в структуру объединенной Европы, выбравшей английский язык в качестве языка межнационального общения, интерес к образовательным стандартам, принятым Советом Европы и применяемым во всем мире, определяет необходимость создания специализированных систем обработки многоязычной информации, в частности, систем компьютерной поддержки обучения, в том числе дистантного, а также систем англо-русского машинного перевода методических и специальных текстов в помощь преподавателю.

Многолетний опыт использования компьютерных технологий для решения различных типов задач показал, что получаемый результат во многом зависит от того, насколько правильно произведена автоматическая переработка текста (АПТ) на естественном языке (ЕЯ). Информационные технологии в области естественного языка (лингвистические технологии), реализующие АПТ, являются необходимым условием решения многих задач, относящихся к информационным технологиям в целом.

Новые потребности общества приводят к необходимости сделать информацию активной, то есть обеспечить максимальное использование информации на электронных носителях и содействовать распространению знаний. Это значит, что информацию, получаемую по сетям Интернет или им подобным, нужно уметь искать, классифицировать, уметь извлекать из нее необходимые сведения, составлять резюме и аннотации, хранить их в специализированных (ориентированных на конкретные нужды) информационно-поисковых системах, переводить на другой язык.

Реалии новой Европы требуют специальных средств поддержки совместной деятельности в условиях многоязычной коммуникации, что, в свою очередь, предъявляет особые требования к языковым интерфейсам, к системам ввода/вывода информации на языке конкретного пользователя независимо от языка источника, к системам, позволяющим работать без использования клавиатуры.

Эти же условия требуют обеспечения полной многоязычности информации на

всех этапах ее существования, что может быть обеспечено за счет создания систем генерации и поддержки многоязычной информации, локализации данных и программного обеспечения, за счет создания практических систем автоматического (машинного) перевода и компьютерных обучающих систем.

Рабочее информационное пространство и его структура

В современном мире в условиях открытой и многоязычной научной коммуникации и развития средств непрерывного и открытого обучения возникает целый ряд задач, решение которых связано с качеством и практической применимостью различных информационных технологий, связанных с анализом текстов на естественном языке и звучащей речи. К таким задачам в самом общем виде относятся:

- автоматический поиск, извлечение и обогащение информации и знаний, получаемых из различных мультимедийных, многоязычных источников и источников, связанных с коммуникацией различных участников;
- межъязыковое или многоязычное извлечение, презентация и распространение информации;
- автоматическое обнаружение и «отслеживание» возникающих тем и проблем из неструктуризированных мультимедийных данных;
- использование источников знаний для того, чтобы облегчить разметку знаний и доступ к ним (в качестве таких структурированных источников знаний могут выступать одно- и многоязычные лексиконы, толковые и энциклопедические словари, тезаурусы, энциклопедии и т. д.);
- поддержание вопросно-ответного взаимодействия человека и компьютера или людей с помощью компьютера как посредника для извлечения знаний из источников различной природы, структуры и состава;
- поддержание дистантного обучения в системах открытого образования, включая автоматизированное тестирование уровня знаний, разработку электронных учебников и диалоговых обучающих систем;
- создание интеллектуальных средств для поддержки автоматизированного ведения библиографической работы, анализа и понимания документов для того, чтобы обеспечить возможность доступа к информации различных экспертов или групп экспертов;
- моделирование знаний, надежд, планов, потребностей и намерений пользователей на основе анализа их запросов к различным системам, созданных ими продуктов и взаимодействия с компьютером;
- обеспечение возможности устного диалога с компьютером, поддержки анализа и порождения звучащей речи[1].

Для решения каждой их этих задач в отдельности или их совокупности необходима разработка не только структуры соответствующего информационного пространства, но и средств реализации конкретных функций системы его поддержки и использования, т. е. специализированного автоматизированного рабочего места (АРМ). Выбранные в каждом конкретном случае функции АРМ определяют наполнение такой системы поддержки от «простых» терминологических ресурсов (доступа к всевозможным одно- и многоязычным учебным пособиям, обучающим системам, словарям и глоссариям, находящимся как в памяти компьютера, так и в сети) до более сложных систем. К таким системам относятся системы поиска и обработки информации, машинного перевода, электронные учебники и учебно-методические комплексы, системы специализированных грамматик для предметных областей и т. п.

Кроме того, АРМ должен включать системы представления данных (инструментальные средства) для возможного издания подготовленной информации.

Полученные в рамках многолетних исследований результаты, а также почти сорокалетний опыт работы по автоматической переработке текстов и, в частности, по машинному переводу (МП), в СССР, США, Западной Европе, Японии и Китае[2] дают возможность сформулировать основные теоретические, технологические и прагматические принципы, на основе которых следует строить современные практические и одновременно «интеллектуальные» системы АПТ — лингвистические автоматы.

Следует иметь в виду, что основная идеология создания лингвистических автоматов (ЛА), способных решать перечисленные выше задачи, относится к периоду универсальных компьютеров и, следовательно, всегда рассматривалась возможность функционирования такого автомата в рамках специализированной системы переработки текстовой информации. Повсеместный переход на использование персональных компьютеров потребовал особого подхода к ЛА, поскольку персональный компьютер дает возможность конкретному пользователю организовать свой собственный «информационный центр», которому должны быть переданы специфические функции обработки текста. При этом пользователь должен представлять себе как особенности и возможности функционирования ЛА в целом, так и условия реализации конкретных функций. Однако базовые требования, предъявляемые к ЛА[3], оказались универсальными и не зависящими от реализации ЛА на конкретной модели компьютера.

Лингвистический автомат представляет собой иерархическую систему программных модулей, каждый из которых осуществляет конкретную операцию по переработке текста и может функционировать как независимо, так и в комплексе с другими. Эти модули (подсистемы ЛА) предназначены для выполнения следующих «интеллектуальных» операций:

- ***опознание языка,*** на котором написан конкретный текст, в потоке информации (такое опознавание может осуществляться как на основе заранее сформулированного и заданного в системе списка языков, так и в ситуации теоретически бесконечного набора возможных языков, из которого выбираются тексты на заранее определенных пользователем языках). Эта функция может реализоваться как в письменном тексте, так и в звучащей речи;
- ***индексирование текста,*** которое сводится к распознаванию его основного содержания и, следовательно, тематики: этот процесс может осуществляться как по отношению к заранее определенному списку тематических областей, так и в виде получения лексического индекса (набора ключевых слов), определяющего основную тему текста;
- ***информационный поиск,*** в задачу которого входит извлечение информации по запросу пользователя из базы данных; запросы эти могут быть постоянными или могут меняться в зависимости от потребностей в информации конкретного человека;
- ***реферирование (аннотирование) текста,*** которое заключается в сжатии текста (с задаваемым пользователем коэффициентом сжатия) до набора наиболее информативных предложений, а также в выполнении простейших экспертных функций;
- ***машинный перевод (МП) текстов,*** который может осуществляться в двух версиях. Во-первых, — в рамках АРМ переводчика для получения высококачественного, жестко ориентированного на предметную область, задачи пользователя и тип документации перевода. В этом случае переводчик выполняет, скорее, функции редактора. Во-вторых, подсистема может быть инструментом пользователя, не знающего иностранного языка, который может очень быстро и с небольшими затратами получить приблизительный (грубый) перевод текстов в интересующей его области знаний, перевод, достаточный для понимания информации, передаваемой текстом на иностранном языке;

- ***распознавание и синтез звучащей речи,*** которые обеспечиваются разработкой систем устного ввода и вывода, являющихся необходимым интерфейсом для практических экспертных и справочных систем;
- ***автоматизированное создание словарей,*** которое заключается в компьютерной поддержке лексиконов (резидентных, автоматических и специализированных словарей) для различных областей знаний; сюда же можно отнести выполнение ставших уже рутинными задач по подготовке частотных, алфавитных, обратных словников, списков машинных оборотов, конкордансов и т. п.
- ***автоматизированное создание документов*** определенной структуры и содержания, осуществляемое с помощью компьютера на формализованном варианте естественного языка;
- ***поддержка обучения родному (иностранному) языку*** с помощью компьютера[4]. К этому направлению относятся задачи создания контролирующих, тестирующих и собственно обучающих лингвистических автоматов. Необходимо отметить, что именно эти задачи являются самыми сложными, поскольку требуют моделирования спонтанного диалога на естественном языке, что, в свою очередь, определяет необходимость создания чрезвычайно сложных систем анализа и понимания естественного языка. Отсюда неизбежно следует вывод о том, что сегодня следует определить те функции и виды деятельности, которые действительно целесообразно моделировать с помощью компьютерных обучающих систем.

При построении ЛА должны соблюдаться определенные лингвистические и кибернетические принципы:

- *человеко-машинный принцип функционирования ЛА,* сущность которого состоит в том, что ЛА в идеале должен полностью моделировать лингвистическое поведение человека. Достижение этого уровня функционирования автомата является тем пределом, к которому стремятся при проектировании ЛА. При разработке и модификации ЛА ему должны передаваться различные семантические функции человеческого разума последовательно — от более простых к более сложным. Соблюдение этого принципа определяет необходимость следования второму принципу разработки ЛА;
- *принцип модульности архитектуры ЛА,* реализация которого означает, что любой вариант ЛА представляет собой композицию простых блоков-модулей. При таком подходе обеспечивается совместимость версий автомата, различных как по решаемым задачам, так и по уровням их конструктивной и функциональной сложности. Соблюдение этого принципа позволяет при необходимости исключать какие-либо модули или включать дополнительные.

В условиях решения задач разработки современных технологий профессиональной подготовки в системе открытого образования построение модульного лингвистического автомата обеспечивает возможность получения результата на любом уровне анализа;

- *принцип открытости системы ЛА,* связанный с первыми двумя перечисленными, заключается в том, что не только вся система может развиваться за счет подключения новых модулей, но и сами модули могут неограниченно расширяться и обновляться.

Опыт разработки и внедрения частных систем, включаемых в структуру ЛА, позволил определить три основных принципа, связывающих технологию конкретной разработки с общим подходом к проектированию ЛА:

1) модульно-иерархическая организация всех систем и модулей,

2) разделение базовых модулей, проблемно-ориентированных модулей и модулей, ориентированных на особенности текстов, в лингвистическом и программном обеспечении,

3) использование трансфера как основного способа преобразования текстов.

Основная идея такого лингвистического подхода состоит в выделении групп взаимосвязанных подпроцессов в общем процессе АПТ. Это выделение должно осуществляться так, чтобы взаимодействие подпроцессов обеспечивало определенную устойчивость системы к различным входным данным и одновременно позволяло сохранить открытую модульную структуру. Изменение набора модулей при решении различных лингвистических задач и ориентации ЛА на конкретные цели позволяет обеспечивать общую гибкость системы на основе сохранения ее общей сердцевины. Поэтому при проектировании и развитии ЛА особое значение приобретает конструирование универсальных блоков-модулей.

Говоря о лингвистическом, и отчасти программном обеспечении ЛА, следует подчеркнуть, что оно должно учитывать:

- потребности пользователей системы (экспресс-информация, справочная информация, извлечение фактографической информации по запросу, сигнальный или высококачественный перевод с постредактированием, обучение языку, тестирование, управление обучающей последовательностью и т. д.);
- особенности информационного потока (объем и типы отдельных текстов и их потоков, возможности «дружественной» коммуникации, типы входных языков, возможности пре-, интер- или постредактирования);
- специфику терминологии и грамматики текстов конкретной предметной области (ПО);
- типологию обрабатываемых языков.

Центральным модулем ЛА, как, впрочем, и любой другой системы АПТ, является лингвистическая информационная база (ЛИБ) данных и знаний, которая должна содержать формализованное описание фонемного строя, лексики, морфологии и синтаксиса входного и выходного языков, а также семантики обрабатываемой предметной области. Кроме того, должен быть обеспечен интерфейс между ЛИБ и программным обеспечением, реализующим анализ входного и генерацию выходного текстов или звучащей речи. Исходя из этих задач, ЛИБ образуется из следующих составляющих:

- блоки лингвистической компетенции (ЛК) на уровне входных языков, включающих описания фонемного состава, лексики и грамматики этих языков, а также процедур анализа;
- блоки ЛК на уровне выходного языка, включая описание его лексики и грамматики, а также синтезирующих процедур;
- тезаурус, в котором кумулируется экстралингвистическая информация о ПО;
- грамматические таблицы, реализующие интерфейс между автоматическим словарем как хранилищем лексической и грамматической информации и программными модулями, которые осуществляют процедуры анализа, трансфера и синтеза.

Организация ЛИБ, как и всего ЛА, предусматривает модульное построение. ЛИБ реализуется в виде набора нежестко связанных блоков. Эта модульность позволяет компоновать ЛИБ по мере готовности отдельных блоков и устранять дублирование информации, обеспечивая при этом возможность последовательного, поэтапного решения задач АПТ.

В настоящее время можно говорить о том, что различные версии ЛА реализованы с той или иной степенью полноты. Так, например, реализован ЛА для сквозного понимания документов[5], конечной целью которого является извлечение необходимой информации на английском языке из многоязычного потока документов. В соответствии с этой задачей в ЛА предусмотрены функции:

– оптического распознавания (при вводе документов со сканера) и опознания языка;

– индексации, предполагающей «приклеивание» к тексту ярлыка, содержащего информацию о его источнике, дате получения и названии, и функции хранения в базе данных;

– определения тематики и дальнейшего распространения документа в соответствии с запросами пользователей;

– полностью автоматизированного машинного перевода;

– накопления машинных переводов для дальнейшего использования различными пользователями системы;

– информационного анализа, заключающегося в создании таблицы, в которой в соответствии с запросами на английском языке содержатся элементы информации из многоязычных источников[6].

Поскольку функции ЛА в той или иной степени реализованы в средствах поиска, предоставляемых пользователям Интернета и других сетей, вопрос об особенностях формирования ЛА для конкретных задач оказывается чрезвычайно важным не только с теоретической точки зрения. Он важен и с точки зрения конкретных пользователей: что им нужно знать об особенностях автоматической переработки документов для того, чтобы получить необходимую информацию с максимальной скоростью, полнотой и точностью.

Любая обучающая диалоговая человеко-машинная система может рассматриваться как вариант ЛА. При этом нужно иметь в виду, что, так же как развитие инженерной лингвистики повлекло за собой обогащение идей и методов не только в прикладной, но и в теоретической лингвистике, и разработка обучающих ЛА приводит к дальнейшему развитию и оптимизации методов преподавания языков. Следует отметить, что именно для преподавания языка критическим является наличие/отсутствие систем анализа и синтеза звучащей речи.

Информационное пространство систем обучения языку

Поддержка обучения родному (иностранному) языку с помощью компьютера представляет собой задачу особой сложности. Дело в том, что использование компьютера при обучении языку, как, впрочем, и любому другому предмету, предполагает поддержку полноценного диалога «компьютер—обучаемый».

Следовательно, полноправным участником речевого информационно-дидактического процесса ЛА сможет стать тогда, когда подсистема поддержки обучения родному (иностранному) языку будет включать следующие блоки-модули:

- блок описания модели внешнего мира на том уровне, который необходим для решения конкретных дидактических задач;
- блок распознавания ситуации, который должен устанавливать контакт с обучаемым и определять уровень его знаний;
- блок принятия решения, в котором устанавливается последовательность обучения в каждом конкретном случае;
- блок обучения, осуществляющий саму дидактическую процедуру, выбранную на основе работы всех предыдущих модулей, и включающий в качестве обязательного компонента подблок управления обучением;
- блок тестирования и диагностики, определяющий уровень знаний и скорость обучения в каждом конкретном случае, и т. д.;
- блок системно-нормативного анализа входного текста, предназначенный для «понимания» входного текста, т. е. для анализа его грамматической и семантической структуры, для установления степени корректности по отношению к требуемой реакции.

Исходным модулем при построении любой подсистемы ЛА является описание внешнего мира. Но если в других случаях такая модель может ограничиваться ее отражением в виде автоматического словаря или тезауруса, в случае создания подсистемы поддержки обучения языку с теоретической точки зрения нам необходимы три частные модели: модели знания, модели причинно-следственных связей и модели возможных решений.

– *Модель знания* определяется предметной областью и реализуется в виде автоматического словаря, отражающего эту ПО, или тезауруса. Чаще всего применяется древовидный тезаурус, в котором с помощью дескрипторов описываются, а затем с помощью родо-видовых связей классифицируются основные понятия семантического пространства. При этом связи между элементами могут фиксироваться либо непосредственными отсылками между словарными статьями, либо лексико-семантическими описаниями, комбинаторика которых отражает структуры связей в ПО[7].

– *Модель причинно-следственных связей* задается в виде семантической сети и совмещается со статической тезаурусной моделью ПО. В идеале семантическая сеть должна рассматриваться как список возможных предикатов (потенциально возможных действий) и их аргументов. В то же время семантическая сеть представляет собой модель реляторов, выражающих более тонкие (по сравнению с фиксируемыми в тазаурусе) ассоциативные связи.

– *Модель возможных решений* строится, как правило, в виде набора типовых ситуаций — фреймов. Фрейм является инструментом описания структуры ситуации с помощью семантической сети. Нетерминальные узлы сети содержат общие сведения (характеристики), остающиеся неизменными при любых вариантах ситуации, описываемой фреймом, а терминальные узлы заполняются конкретными и частными сведениями.

Типовыми фреймами, которые должны быть разработаны при организации обучающего ЛА, являются:

- фрейм построения вопросов, который задает ЛА;
- фрейм анализа ответа обучаемого;
- фрейм оценки ответа обучаемого;
- фрейм тестирования;
- фрейм определения уровня обучаемого и/или истории его обучения.

Тогда, например, для фрейма анализа ответа обучаемого нетерминальными узлами будут ситуации получения ответа, что требует предварительной настройки либо на потенциально возможные структуры ответа, либо на выбор варианта из меню, либо на процедуру отождествления получаемого ответа с потенциально возможными и предусмотренными в системе реакциями и т. п.

В терминальных узлах фреймов накапливается информация о сделанных ошибках (за один сеанс обучения или за весь цикл), о типах вопросов и т. д.

Фреймы должны быть организованы в иерархическую последовательность так, чтобы система могла по значениям терминальных узлов (по количеству ошибок, времени ответа, типам ошибок и т. п.) автоматически подключать следующий фрейм и формировать информацию о процессе обучения конкретного человека (количество ошибок, сделанных на конкретную тему, общая предварительная оценка, тематика работы, уровень знаний). Использование понятия фрейма является удобным операционным средством для описания того, что нужно формализовать в подсистеме поддержки обучения родному (иностранному) языку.

Для создания систем компьютерного обучения целесообразно следовать общепринятой типологии фреймов на статические фреймы-сценарии и динамические фреймы-планы. При этом сценарий строится как статическая структура, в которой задано название (метка) ситуации, определена причина ее возникновения и a priori задан набор и последовательность сцен. В качестве набора сцен можно рассматривать последовательность макродействий обучающего ЛА, которая определяется, с одной стороны, обучающей последовательностью, а с другой — особенностями работы обучаемого.

С помощью фрейма-плана имплицируются причинно-следственные связи между конкретными ситуациями, возникающими при использовании обучающего ЛА. Сценарий и план различаются по

ряду параметров: сценарий строится по принципу предсказуемости, а планы описывают множество выборов, которые определяют путь достижения конкретных целей. Можно, вероятно, считать, что обучающая последовательность, заданная в обучающем автомате, представляет собой фрейм-план, а реализуется она в виде мобильной системы фреймов-сценариев. При этом выбор конкретного сценария определяется особенностями протекания процесса обучения в каждом случае. Таким образом, моделирование целей и решений в обучающем ЛА не должно ограничиваться заданием статических сценариев-фреймов.

Создание таких технологий требует от их разработчиков детального моделирования процесса обучения и выработки принципов классификации всех его составляющих и их разделения по меньшей мере на три группы.

Во-первых, должны быть определены те составляющие процесса обучения, которые при всех условиях и технических возможностях остаются в ведении педагога.

Во-вторых, необходимо уметь определять компоненты стратегии научения, которые целесообразно моделировать с помощью компьютера. При этом столь же важно определить требования к парку компьютеров, с помощью которого можно работать.

И, наконец, нужно выделить те элементы, которые требуют применения различных мультимедийных средств, и определить их структуру и состав.

В целом это очень сложная задача, которая требует совместной творческой работы специалистов в различных областях знаний, педагогов, психологов, лингвистов и программистов, методистов и специалистов по передаче информации. Следовательно, новые образовательные технологии — это то, что должно создаваться на основе единой концепции, единых принципов и подходов, которые могут быть выработаны только совместными усилиями.

Принципы организации автоматического рабочего места филолога

Средства образовательных технологий должны быть организованы в систему, прообразом которой может быть автоматизированное рабочее место (workstation). Идея создания автоматизированного рабочего места (АРМ), предназначенного для различных групп пользователей[8, 9], возникла еще в 80-е годы прошлого столетия.

Такое рабочее место в то время рассматривалось как терминал для автоматической переработки текста, снабженный экраном и клавиатурой и имеющий связь в режиме разделения времени с универсальной машиной[10]. Сам термин АРМ (translator's workstation) был предложен А. Мэлби в 1981 году[11]. С его точки зрения, АРМ должен представлять собой микрокомпьютер с двумя дисководами для гибких дисков, дисплей, клавиатуру, небольшой принтер и порт для связи с универсальной машиной. Такое устройство позволило бы переводчику осуществлять оперативное редактирование перевода и поиск необходимой информации в словарной базе. С переходом на персональные компьютеры идея АРМ стала реальностью[12].

Создание АРМ, таким образом, предусматривает как решение традиционных задач, относящихся к задачам лингвистического автомата или лингвистического процессора, так и активное взаимодействие с сетью Интернет и информацией в ней. Влияние сети Интернет оказывается решающим для создания АРМ, поскольку в этой сети созданы возможности для широкого обмена многоязычной и мультимедийной информацией и, следовательно, возникает осознанная необходимость объединения переводческих ресурсов сети.

Для обеспечения доступа к мультимедийной информации в АРМ необходимы специальные средства, объединяющие обработку изображения, аудио- и видеоряда, звучащей и письменной речи, поиск

и извлечение информации (фактографический анализ) из различных баз знаний и реальных потоков информации, реферирование, перевод, а также проектирование представления результатов анализа, что тоже должно включать в себя мультимедийные средства[13].

Сегодня можно рассматривать АРМ филолога как вид лингвистического автомата, в котором реализованы возможности для работы:

– профессионального переводчика;
– лингвиста-исследователя;
– лексикографа;
– литературоведа;
– студента-филолога;
– редактора[14].

Соответственно, функции АРМ определяют выбор конкретной системы поддержки от «простых» терминологических ресурсов (доступа к всевозможным одно- и многоязычным словарям и глоссариям, находящимся как в памяти компьютера, так и в сети) до более сложных, таких как системы машинного перевода, системы специализированных грамматик для предметных областей и т. п. Кроме того, АРМ должен включать системы представления данных (инструментальные средства) для дальнейшего издания.

Анализ различных АРМ и отдельных систем обработки и перевода текстов, разнообразие словарей, тезаурусов и глоссариев показывает, что в рамках педагогики и методики в целом, а также в области лингвистики, литературоведения и методики преподавания иностранных языков, не существует системы, ориентированной на потребности преподавателей различных дисциплин и, в частности филологов, педагогов и методистов. В то же время разнообразие и частое несоответствие терминологических систем в разных языках и методических концепций в разных странах приводят к тому, что часть лингвистического (филологического) сообщества оказывается не в состоянии адекватно оценить уровень разработки конкретной проблемы.

Исходя из этого, можно утверждать, что необходимо выработать теоретические основы и практические методы создания автоматизированного рабочего места педагога, позволяющего ему выполнять как конкретные исследовательские и учебные задачи (исследование текстов, записанных в электронном формате, анализ и перевод текстов, выполнение учебных заданий, разработка тестовых баз данных и т. д.), так и проводить самостоятельные научные исследования на базе больших массивов текстов.

Автоматизированное рабочее место в системе самообразования должно включать:

– комплекс электронных учебников и обучающих программ по конкретной предметной области (или иерархию таких учебников и обучающих курсов по конкретному направлению обучения), а также

– комплекс средств, обеспечивающих активное взаимодействие с сетью Интернет и информацией в ней.

Влияние сети Интернет оказывается решающим для создания АРМ, поскольку в этой сети созданы возможности для широкого обмена многоязычной и мультимедийной информацией и, следовательно, возникает осознанная необходимость объединения ресурсов сети. Для этого объединения необходимы специальные системы поиска информации, обеспечивающие возможность гибкого соотнесения неполно и нечетко выраженных запросов пользователя на получение информации, которой у него нет, с массивами текстов, которые ему неизвестны. Поскольку основная масса информации в сети является текстовой, для ее поиска и тематической обработки необходимы специальные средства обработки естественного языка — языковой интерфейс.

Для обеспечения доступа к информации, хранящейся в сети Интернет, необходимы специальные средства, объединяющие обработку изображения, аудио- и видеоряда, звучащей и письменной

речи, поиск и извлечение информации (фактографический анализ) из различных баз знаний и реальных потоков информации, реферирование, перевод, а также представления результатов анализа, что должно включать в себя мультимедийные средства

Информационно-поисковые системы (ИПС) могут создаваться в двух вариантах: как системы избирательного распространения информации, в которых поток документов анализируется относительно фиксированного набора запросов, и как системы поиска информации в массиве документов, когда запрос может формулироваться каждым пользователем системы при обращении к ней. Таким образом, ИПС обрабатывают информационные запросы, идентифицируют и извлекают соответствующие единицы из массива документов.

В системе проектирование на основе WWW и Gopher возможен автоматизированный доступ к различным массивам информации и базам данных. Соответственно в базах данных конкретного АРМ могут храниться полные тексты документов на разных языках, включая изображения и графики. Это могут быть полнотекстовые базы (например, Newsbyte, News Service, Business Wire и т. д.), тексты аннотаций (ср. базу данных Cambridge Scientific Abstracts), а также описания этих документов в виде наборов ключевых слов и словосочетаний. Автоматизация составления таких наборов (поисковых образов документов) может выполняться предварительно на уровне рубрикации текстов и происходит при введении текста в систему, однако точность и полнота такого индексирования зависят от задач системы информационного поиска.

Таким образом, создание автоматизированного рабочего места, обеспечивающего специалиста возможностями поиска и обработки информации на естественном языке, представляет собой актуальную задачу, решение которой потребует объединения усилий специалистов в различных областях знаний. Решение этой задачи можно начинать с создания модельного АРМ в той области, где накоплен большой опыт работы с текстами на естественном языке, — в филологии.

Подобный научно-учебный автоматизированный комплекс «Филология» (автоматизированное рабочее место филолога) должен представлять собой комплекс программных, лингвистических и технических средств, обеспечивающих удобство работы и потребности как студента, обучающегося по направлениям филологического цикла, так и преподавателей и исследователей в различных отраслях этой области знаний (лингвистов, литературоведов, методистов и т. д.).

Создание АРМ филолога предусматривает как решение традиционных задач, решаемых в рамках системы открытого образования, так и активное взаимодействие с сетью Интернет и информацией в ней.

Выбранные в каждом конкретном случае функции АРМ определяют выбор конкретной системы поддержки от «простых» терминологических ресурсов (доступа к всевозможным одно- и многоязычным учебным пособиям, обучающим системам, словарям и глоссариям, находящимся как в памяти компьютера, так и в сети) до более сложных, таких как системы поиска и обработки информации, машинного перевода, системы специализированных грамматик для предметных областей, системы тестирования и т. п. Особой частью АРМ должна быть система глоссариев по различным направлениям лингвистических, литературоведческих и методических теорий, включающая не только переводы терминов на различные языки, но и их толкование в авторитетных изданиях.

Кроме того, АРМ должен включать системы представления данных (инструментальные средства) для возможного издания подготовленной информации.

Таким образом, в структуру комплекса, который следует разрабатывать в рамках

конкретной специальности, должны войти программные и лингвистические модули:

- лингвистический автомат как комплекс средств АПТ (функции компрессии информации, поиска информации, перевода научных и учебных текстов и т. д.). Эта часть комплекса может непосредственно использоваться в учебном процессе при обучении переводу, редактированию, аннотированию текстов, в письменной практике и т. д.;
- база полнотекстовых данных, предполагающая хранение, модификацию и поиск текстов произведений художественной и научной литературы на разных языках с формированием массивов параллельных и псевдопараллельных текстов. Эта часть комплекса может непосредственно использоваться в учебном процессе для анализа конкретных лингвистических и литературоведческих фактов, для проведения сравнительного стилистического анализа, для изучения особенностей авторского стиля и т. д. Кроме того, подобная база является важным источником сведений для создания словарей разного состава и назначения;
- база терминологических данных, предполагающая хранение, структуризацию, составление тезауруса терминов и поиск терминов по предметной области «Филология». Эта база может быть использована непосредственно в учебных целях для самостоятельной работы студентов в рамках курсов по общему языкознанию, литературоведению, лексикологии, теоретической грамматике, стилистике и другим базовым теоретическим и прикладным дисциплинам;
- база референтных и автоматических словарей, словарей на машинных носителях, учебных словарей, объединенных в единый комплекс, позволяющий в рамках АРМ использовать любую накопленную словарную информацию в учебных и научных целях;
- база специализированных лингвистических программных средств, обеспечивающих получение словарей разных видов (алфавитных, частотных, обратных, конкордансов и т. д.) на основе информации из других баз данных. Работа с прикладными лингвистическими программами позволит студентам овладеть соответствующими научными методами, специалисты-филологи получат возможности исследования больших массивов текстов;
- постоянно пополняемая библиографическая база данных по филологии, позволяющая студентам и преподавателям производить библиографический поиск;
- база обучающих программ и процедур тестирования, предназначенная для студентов, изучающих иностранные языки и русский язык;
- база средств обработки мультимедийной информации, поддерживающая обучение различным аспектам межкультурной коммуникации и речевого поведения.

Разработка подобного проекта является долгосрочной программой, в рамках которой следует проводить исследования, позволяющие методом пошагового приближения дать возможность достижения основной цели. Понятно, что для своего полного воплощения такой проект требует привлечения большого коллектива исполнителей, сложной техники и финансирования, поэтому общая концепция проекта рассматривается как цель, к которой возможно последовательное приближение.

На данном этапе задачей исследований является определение иерархии задач и последовательности их решения, а также

– выбор и обоснование типа системы управления базой данных для создания баз на основе имеющихся текстовых вариантов лингвистических тезаурусов и учебных терминологических словарей;

– создание базовой модели системы машинного перевода текстов по филологии с английского языка на русский на основе разработанной в лаборатории машинного перевода системы СИЛОД (версия WORD+);

– создание базы учебных модулей, модифицируемых и развиваемых студентами.

Использование новых возможностей, которые предоставляют нам информационные технологии, с одной стороны, позволяет получить результаты, которые были невозможны ранее, с другой — определяет новые направления исследований.

* * *

Сегодня очень важным остается вопрос о том, как должно создаваться информационное пространство и как подобная система должна встраиваться в общую стратегию обучения. Ясно, что нет ни возможности, ни необходимости передавать все функции филолога и преподавателя компьютеру. Период, когда нужно было доказывать, что компьютер способен заменить преподавателя в отдельных аспектах его работы, прошел.

Говоря о современной компьютерной лингводидактике, т. е. об обучении языку с помощью компьютера (ср. английский термин CALL — computer-assistant language learning), мы видим необходимость в разработке специализированных учебных курсов, в которых программные средства применялись бы только там, где это необходимо с методической точки зрения[15]. Более того, в качестве таких программных средств можно использовать компоненты ЛА. Так, например, можно говорить об использовании системы МП при обучении переводу и редактированию как видам деятельности; систему аннотирования можно использовать при обучении компрессии текста; работа с автоматическими словарями полезна как часть практической работы по курсу лексикологии и т. д.

Особая ситуация в системах компьютерной поддержки обучения и исследования возникает с привлечением сети Интернет, которая при дистанционном обучении может использоваться и как средство распространения учебных материалов, и как средство прямой связи между обучаемым и преподавателем. Естественно, через Интернет возможен доступ к различным обучающим системам. Опыт организации занятий с использованием Интернета в режиме реального времени показывает, что для них важен учет разницы во времени, необходимо также разработать специальные формы контроля. Кроме того, при организации дистанционного обучения особую роль играет выбранная в системе концепция электронного учебника. Такой концепции, к сожалению, сегодня нет, и можно только утверждать, что без тщательного теоретического анализа этой проблемы создать такую концепцию невозможно, а решения ad hoc в данном случае неприемлемы.

Высоко оценивая возможности как самих компьютерных систем, так и сети Интернет, многократно увеличивающих их возможности, хочется предостеречь от переоценки тех «быстрых» результатов, которые могут быть получены в рамках систем открытого образования. Весь опыт применения компьютеров в гуманитарных областях, особенно в области машинного перевода[16], показывает, что завышенные ожидания, с одной стороны, и неоправданные обещания, с другой, могут привести к катастрофическим последствиям для всего направления в целом.

Следует еще раз подчеркнуть, что направления для развития и использования информационных технологий в области открытого образования должны выбираться не потому, что нам что-то известно и мы умеем этим инструментом пользоваться (или разрабатывать его). Основная задача состоит в том, чтобы очень осторожно и адекватно выбрать те средства и информационные технологии, которые действительно необходимы для реализации учебного процесса.

Разработке новых образовательных технологий необходимо учить, учить пользователей и готовить разработчиков. К сожалению, и специалисты в отдельных областях знаний, и лингвисты, и все педагогическое сообщество не готовы не только решать эти задачи, а даже и участвовать в их решении. Проблема кадров оказывается не менее важной, чем проблема создания концепции. Если говорить о филологии, то, к сожалению, сегодня в практически применяемых информационных технологиях, связанных с обработкой текстов на ЕЯ, очень многое (особенно в обучающих системах) решается инженерами, программистами, математиками, но не филологами или методистами.

Следовательно, есть настоятельная необходимость разработки и внедрения специальных образовательных программ, направленных на подготовку специалистов одновременно в области лингвистики, методики и компьютерных информационных технологий.

На филологическом факультете уже существует профиль для подготовки бакалавра «Иностранный язык и компьютерная лингводидактика», магистерская программа «Информационные технологии в филологии», и лицензируется подготовка аспирантов по специальности 10. 02. 21 — прикладная и математическая лингвистика. То есть мы пытаемся создать полный курс подготовки специалистов, способных решать задачи создания новых информационных технологий.

Профессионалы, которых мы готовим, — это филологи, обладающие базовыми знаниями в области **математики** как средства систематизации, формализации и обобщения. Соответственно, блок математических дисциплин включает учебные курсы по основам дискретной математики, статистики и теории вероятностей, программирования применительно к лингвистическим задачам, а также теории обучающих систем.

Для специалистов этого класса необходимы знания и умения в области информационных технологий, соответственно, в программу подготовки включены дисциплины, охватывающие:

- проблемы разработки и использования лингвистических автоматов разного уровня,
- рассмотрение особенностей алгоритмизации и экспликации лингвистических и методических задач,
- проблемы автоматизации в лексикографии, особенности применения информационных технологий и формальных методов при лингвистическом и литературном анализе, при создании систем обучения языку.

Кроме того, специалисты по компьютерной лингводидактике должны получить достаточно серьезную традиционную лингвистическую подготовку как в области теоретического языкознания, так и в области английского и русского языков. Соответствующие практические и теоретические курсы включены в разработанную программу бакалавриата в том же объеме, что и для профиля «Иностранный язык и зарубежная литература». Базовое образование в области литературы дается в том же объеме, что и на других профилях.

Подготовка специалистов-бакалавров по иностранному языку и компьютерной лингводидактике предполагает, что кроме овладения всем комплексом знаний, предлагаемым программой, каждый выпускник должен разработать реальный модуль обучающей программы, который станет ядром его дальнейших разработок в профессиональной деятельности в школе. То есть в рамках бакалавриата мы должны выпустить специалиста с «инструментом» для дальнейшей работы. Этот модуль он может развивать дальше в рамках специалитета или магистерской программы. Магистерская программа, в свою очередь, предполагает подготовку специалистов, готовых как к исследова-

тельской работе в области создания новых лингвистических проектов, так и к преподаванию в школе и вузе. Эта программа включает дисциплины как общефилологического направления, так и специализированные, направленные на углубление полученных знаний в области прикладной лингвистики, использования Интернета, баз данных, разработки лингвистических автоматов. Для филфака в этом направлении очень важно подготовить специалистов в области компьютерной лексикографии, создания и ведения баз данных, применения современных технологий в литературоведческом анализе и обучении.

Можно сказать, что описанное направление в обучении — выполнение социального заказа. Но не менее важно подготовить будущих исследователей, которые смогут учесть в своей работе совершенно новые возможности, которые сегодня открыты для филолога.

Следует еще раз подчеркнуть, что направления для развития и использования информационных технологий в области открытого образования должны выбираться не потому, что нам что-то известно и мы умеем этим инструментом пользоваться (или разрабатывать его). Основная трудность, по нашему мнению, состоит в том, чтобы очень осторожно и адекватно выбрать те средства и информационные технологии, которые действительно необходимы для реализации учебного процесса.

Сегодня следует уделить особое внимание определению принципов разработки подсистем поддержки обучения (в том числе родному и иностранному языку) как элементов дружественной обучающей среды, так необходимой всем, кто учится, и кто учит.

ПРИМЕЧАНИЯ

[1] *Maybury M. T.* Intelligent Multimedia Information Retrieval/ Menlo Park: AAAI/MITPRESS, H. T. Keynote. Intell. Multimedia for the new Millenium. Proceedings of Eurospeech'99. Budapest, Sept. 6-0. Vol. 1. 1999. P. KN1–KN15.

[2] *Пиотровский Р. Г.* Лингвистический автомат (в исследовании и непрерывном обучении). СПб., 1999.

[3] *Беляева Л. Н.* Лингвистические автоматы в современных информационных технологиях. СПб., 2001; *Беляева Л. Н.* Проблемы анализа корпусов параллельных текстов с различной графикой // Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных». СПб., 2002. С. 27–39.

[4] *Карамышева Т. В.* Компьютерная лингводидактика. СПб., 2000.

[5] *White J. S., O'Connell T. A.* The ARPA MT Evaluation Met Lessons, and Future Approaches // Proceedings of the 1994 Confer. Machine Translation in the Americas (AMTA). 1996.

[6] Там же.

[7] *Bernardi U., Gieselmann P., McLaughlin S.* A Taste of MALT // Machine Translation in the Information Age. Proc. of MT Summit VIII. Spain, 2001. P. 43–48.

[8] *Arthern P. J.* Machine Translation and Computerized Terminology Systems: A Translator's Viewpoint // Snell D. M. (ed.) Translating and the Computer. Proceedings of a Seminar, London, North-Holland, Amsterdam, 1979. P. 77–108.

[9] *Melby A. K.* Translators and Machines — Can they Cooperate? Meta, 26. 1981. P. 23–34.

[10] *Arthern P. J.* Указ. соч. С. 95.

[11] *Melby A. K.* Указ. соч. С. 219.

[12] *Bernardi U., Gieselmann P., McLaughlin S.* Указ. соч.

[13] *Maybury M. T.* Указ. соч.

[14] *Hutchins J.* The Origins of the Translator's Workstation. Machine Translation, 13.1998. P. 287–307.

[15] *Карамышева Т. В.* Указ. соч.

[16] *Hutchins J.* (ed.) Early Years in Machine Translation // Series III. Studies in the History of the Language Science. Vol. 97. Amsterdam: John Benjamins Publishing Company. 2000.

L. Beliaeva

INFORMATION SPACE AND ITS STRUCTURE IN THE FRAMEWORK OF MODERN PHILOLOGICAL EDUCATION

Modern specialist in the domain of philology needs not only knowledge and abilities in the information technology area, but specialized tools for supporting his/her research and methodical studies. Such tools are to provide a philologist with special means for both solving particular research and educational tasks (analysis and text translation, variety of electronic text studies, different tasks in language teaching and learning), and for conducting independent research on the large text files in the framework of corpus-based linguistics. All our experience shows that such information support necessitate creation of a specialized philological workstation (WS). Such WSs are to be implemented as a complex of software, hardware and linguaware which provides for convenient work both for students, training in philology and language teaching, and educational linguists and researchers in various branches of this knowledge domain (linguists, specialist in literature, etc).