2002 | Том 5 | Выпуск 2

Лингвистическое обеспечение электронных библиотек

А.Б. Антопольский

НТЦ "Информрегистр"

В настоящей статье делается попытка очертить круг проблем, связанных с созданием и использованием языковых средств электронных библиотек.

Актуальность данной задачи не вызывает сомнений. Бурное развитие информатики в последние годы, в том числе невероятно быстрое развитие Интернета породили массовое создание электронных информационных ресурсов по всем областям знания и в самых различных видах деятельности. Специалистам очевидна необходимость организации этих ресурсов в систематически организованные массивы, в которых было бы удобно и эффективно проводить информационный поиск, в том числе неподготовленному пользователю. Очевидно также, что для этой цели нужны адекватные языковые средства.

Интернет стал практически единой средой для создания и размещения информационных ресурсов и для поиска в них. При этом, с одной стороны, он предоставляет множество возможностей, с другой стороны, накладывает определенные ограничения на выбор и применение языковых средств. Класс информационных технологий, направленный на организацию электронных ресурсов в условиях глобальных сетей и организации поиска в них, стал называться *электронными* или *цифровыми библиотеками.* Проблематика электронных библиотек (ЭБ) в целом достаточно полно изложена в работе [1]. Обсуждение понятия ЭБ и связанных с ним проблем имеется также в работе [2].

В то же время ощущается недостаток исследований, обобщающих опыт работ по созданию языковых средств, и ориентирующих разработчиков при выборе этих средств для ЭБ.

Характеристика состояния проблемы. Обсуждение поставленных в статье проблем выявило сложную ситуацию в данной области информационных технологий, обусловленную спецификой современного периода развития отечественной информатики.

Дело в том, что последние 10-12 лет явились периодом кризиса отечественной науки вообще и информатики в частности. В течение 1970-1980-х гг. исследования в области информационных систем и информационного поиска велись широким фронтом, и появлялось множество публикаций, посвященных этим проблемам и, в

частности, лингвистическому обеспечению. Однако примерно с 1990 г. их число резко сократилось, а обобщающих монографий просто не было.

Отчасти это обусловлено тем, что основные исследования и разработки этого направления информатики в СССР были сосредоточены в сети органов научно-технической информации, которая практически рухнула в результате всех событий 1990-х гг. Другой причиной была резкая смена поколений информационных систем: сначала переход с больших ЭВМ на персональные, а затем распространение Интернета. В результате к концу века в стране практически исчезли созданные в 1980-х гг и ранее информационные системы, основанные на известных моделях лингвистического обеспечения.

В то же время 1990-е гг. стали десятилетием рождения нового класса АИС – таких как поисковые машины Интернета, или распространение БД вместе с поисковыми средствами на компакт-дисках (например, с правовой информацией). В этих новых системах по многим причинам, в том числе субъективным, почти не использовался опыт классических ИПС прошлых лет.

В результате возник разрыв между достижениями отечественной информатики прошлых десятилетий и современным состоянием теории и практики поисковых машин и ЭБ.

Автор занимается этой проблемой, можно сказать, с самого начала, с 1960-х гг. и имел возможность воочию наблюдать все этапы эволюции этого направления информатики, как его подъемы, так и падения. Поэтому представляется очень важным хотя бы частично заполнить разрыв между поколениями исследователей и изложить основы создания и использования лингвистического обеспечения информационных систем с учетом как прошлых достижений, так и сегодняшнего состояния в данной области.

Теоретические основы лингвистического обеспечения ЭБ. В настоящее время в информатике отсутствует общая теоретическая модель, охватывающая основные аспекты речевого общения (коммуникативного взаимодействия) человека и информационной системы, в частности ЭБ. В то же время ряд фундаментальных дисциплин изучает коммуникативное взаимодействие в различных аспектах.

Прежде всего, это *общая теория коммуникации*, в том числе теория речевых актов, в рамках которой исследуется структура коммуникативных процессов, их типы, особенности разных типов коммуникативных процессов. Это, безусловно, *лингвистик*а, исследующая языковые формы коммуникативного взаимодействия, *психология*, исследующая формы человеческих реакций в процессе взаимодействия, *логика*, изучающая формальную структуру высказываний и систем классификации, *семиотика*, исследующая структуру знака и функционирование знаковых систем. Важной теоретической базой коммуникации является *теория пресуппозиции*, которая предлагает механизмы экспликации знаний, инвариантных для пользователей в процессе функционирования ЭБ.

Заметим, что при практическом проектировании информационных технологий,

используемых в современных ЭБ, разработчики почти никогда не обращаются к инструментарию базовых теоретических дисциплин. В то же время представляется очень важным использовать эти достижения при проектировании ЭБ, чтобы реже испытывать эффект наступания на грабли. Сделать это можно только в вузе. Поэтому важнейшая задача обучения специалистов по информационным технологиям – дать им хотя бы самые общие представления о теории тех процессов, которыми они практически занимаются.

Информационные технологии, связанные с лингвистическим обеспечением. Создание и использование средств лингвистического обеспечения ЭБ опирается также на многие информационные технологии, влияющие на создание и реализацию языковых средств. Прежде всего, это общие технология функционирования и архитектура информационных систем, это, безусловно, теория информационного поиска, классические и современные решения в области технологий баз данных и, наконец, технология пользовательских интерфейсов, особенно в среде Интернета.

Новой и очень важной для исследований в области лингвистического обеспечения является технология языков разметки. По мнению автора, именно развитие этих языков во многом будет определять перспективы и возможности применения языковых средств в электронных библиотеках, подобно тому, как 20 лет назад возможности поисковых систем определялись теорией и практикой систем управления базами данных.

Необходимо отметить, что теория и практика лингвистического обеспечения ЭБ развивалась в рамках того направления информатики, которое иногда называлось “информационно-поисковыми системами”. Это направление в основном оперировало с текстовыми документами, точнее говоря, с плохо структурированной информацией. Направление информатики, оперирующее с хорощо структурированной информацией, опиралось в основном на теорию и практику баз данных. Теория баз данных и эволюция основных понятий этой теории подробно изложена в фундаментальной работе М.Р.Когаловского [3].

Различные подходы к определению ЛО. В литературе по информатике накопилось много различных подходов к понятию ЛО и соответственно, различных определений этого понятия (или близких понятий “информационно-поисковые языки”, “языковые средства АИС” и др.). Кратко рассмотрим основные подходы.

Наиболее известным является *классический* подход, при котором лингвистическим обеспечением называют комплекс информационно- поисковых языков, прежде всего, классификационных и вербальных (дескрипторных). Этот подход ведет свое начало от классического труда “Основы информатики” [4] и распространен среди разработчиков систем, которые обычно относят к НТИ. С небольшими изменениями этот подход принят и в теории автоматизированных библиотечных систем. В последней, однако, в отличие от классического подхода в понятие ЛО обычно включают и языки библиографических данных.

Существует подход, который можно назвать *“лингвистическим”,* поскольку он органически вытекает из лингвистического взгляда на информационные системы

и который развивают в основном специалисты по прикладной и компьютерной лингвистике. В соответствии с этим подходом лингвистическое обеспечение – это комплекс средств, используемых для автоматической обработки текстов на естественном языке (включая обработку запросов и поиск), т.е. прежде всего, языковые процессоры. Примером является взгляд, излагаемый в работе [5].

Более общим является подход, который следует определить как *“семиотический*”, поскольку он исходит из классических семиотических представлений о языке как системе знаков разного уровня, начиная, естественно, с алфавита. При этом подходе лингвистическое обеспечение определяется как “*средства представления информации в виде данных и интерпретации этих данных”.* Такой взгляд развивал в ряде работ, например, в [6], автор данной статьи. При этом подходе в состав ЛО нужно, например, включать средства кодировки алфавитов или форматы представления данных, но не нужно включать инструментальные языки программирования.

Иногда в литературе можно встретить представление об языковых средствах, которое можно назвать “*программистским”.*

Сторонники такого подхода опираются на полисемию термина “язык”, который, как известно, может обозначать в информационной литературе не только средства представления данных, но и средства манипулирования данными, включая инструментальные средства программирования и другие формальные системы. К тому же средства манипулирования данными в последние годы интегрируются с языками описания данных в рамках языков высокого уровня, которые все ближе к тому, что можно назвать формализованным естественным языком и все дальше от обычного представления от обычных инструментальных программных средств. Таковы, например, языки разметки типа SGML или XML.

При “программистском” взгляде в составе ЛО могут оказаться вообще все языковые средства пользователя, причем несущественно, носят ли они характер языков описания данных, представления данных или манипулирования данными.

Наконец, можно отметить подход, зафиксированный в *нормативных документах по АСУ* (группа ГОСТ 34), в которых разделяются информационное и лингвистическое обеспечение. При этом основной тип ИПЯ этих систем – классификаторы - эти нормативные документы относят к информационному обеспечению, а на долю лингвистического обеспечения остаются только правила оформления естественно-языковых единиц этих классификаторов, т.е. чисто лексикографические аспекты.

Определение ЛО. Представляется что наиболее строгое определение ЛО основано на семиотическом подходе и на понятии ЛО как средств представления данных. Однако опыт автора показал, что реальное распределение функций между постановщиками задач ЭБ, а также разработчиками программного и лингвистического обеспечения таково, что строгое семиотическое определение практически неудобно.

С одной стороны, при строгом определении в понятие ЛО необходимо включать

объекты, которыми традиционно занимаются программисты, такие как системы кодировок, формальные языки запросов или языки разметки. В современных ЭБ к средствам представления данных также относятся языки представления графики, картографии, аудиоинформации, трехмерных и движущихся объектов и других нетекстовых данных. Разработка этих средств также была уделом программистов и далека от интересов разработчиков ЛО ЭБ.

С другой стороны, в область интересов информационных лингвистов (разработчиков ЛО ЭБ), всегда входили не только языковые средства представления данных, но также средства обработки текстов на естественном языке, то есть лингвистические процессоры. Поэтому если попытаться определить ЛО, как объект интересов именно этого класса специалистов, то в него следует включить, во-первых, только семантические средства представления данных, во-вторых, кроме них также лингвистические процессоры, применяемые в ЭБ.

Лингвистические процессоры – это достаточно широкий класс продуктов. В него включают, например, спеллеры, текстовые редакторы, системы морфологического и синтаксического анализа и синтеза текстов, системы автоматического перевода, различные системы компьютерной лексикографии и автоматические словари.

В состав ЛО ЭБ целесообразно включать те процессоры, которые ориентированы на обработку семантических языковых единиц (морфем, слов, словосочетаний), а также высших уровней языка (синтаксиса, сверхфразовых единств).

Предлагаемое ниже определение ЛО не претендует на теоретическую чистоту и рассчитано сугубо на практическое применение.

Лингвистическое обеспечение ЭБ - комплекс языковых средств и процессоров, предназначенных для обработки, представления и поиска письменных текстов на естественном языке, в основном на семантическом уровне.

История разработки ЛО в России. Системная разработка лингвистического обеспечения информационных систем – предшественников ЭБ - велась в России, начиная с 1960-х гг., по нескольким направлениям.

Разработка ЛО ГАСНТИ. Данное направление наиболее полно отражало проблематику создания и использования ЛО, в его реализации участвовали наиболее квалифицированные специалисты бывшего СССР. Проектирование было начато с 1965 г. и в 1969 г. появился первый системный проект ЛО под названием “Комплекс средств индексирования научно-технической информации” [7].

В течение 20 лет шли разработки по широкому классу проблем, связанных с ЛО. В результате к концу 1980-х гг. в ГАСНТИ была сформирована достаточно стройная система языковых средств по всему необходимому спектру функциональных задач АСНТИ, поддержанная развитой системой государственных стандартов и специализированной организационной структурой. В качестве последней выступала Автоматизированная система информационных языков (АСВИЯ),

функционировавшая в ВНИИКИ Госстандарта и в ВИНИТИ РАН. Руководство работами осуществлял ГКНТ, ВИНИТИ, как головная организация, а также коллективные научные органы. Можно утверждать, что созданное ЛО ГАСНТИ соответствовало наиболее высокому уровню информационной науки того времени. Общая модель этого ЛО была зафиксирована в нормативно- правовом документе [8]. Всего в состав ЛО ГАСНТИ входило до 200 тезаурусов и рубрикаторов по всем отраслям народного хозяйства. Были также созданы многочисленные и иногда достаточно мощные лингвистические процессоры. Примером могут служить системы, описанные в работах [5, 9, 10].

Кризис 1990-х гг. в системе НТИ России совпал со сменой поколений АС НТИ (сначала распространение ПЭВМ, затем Интернет), что в совокупности привело к почти полной утрате достижений того времени. В настоящее время из общесистемных языковых средств в ВИНИТИ поддерживается ГРНТИ и частично УДК. По дескрипторным языкам и языкам метаданных системная работа в ГАСНТИ не ведется.

Разработка ЕСКК ТЭИ. Параллельно с ГАСНТИ крупными силами велось создание комплекса языковых средств автоматизированных систем организационно-экономического управления разного уровня, получившего название “Единая система классификации и кодирования технико-экономической информации” (ЕСКК ТЭИ). Чисто научный уровень этих разработок был несколько ниже, чем в ГАСНТИ, зато масштабы работ гораздо больше. В результате была создана система общероссийских классификаторов, число которых к концу 1980–х гг. достигло 35, а их общий объем превысил 3 млн. позиций.

Среди этих классификаторов были такие крупные, как Общесоюзный классификатор продукции (ОКП), предприятий и организаций (ОКПО), объектов административно-территориального деления (СОАТО) и др. Система классификаторов поддерживалась разветвленной службой их ведения, включавшей Главный центр ведения общесоюзных классификаторов при Госстандарте, а также службы ведения в отраслях и регионах. Общее число сотрудников только в этих службах превышало 2 тыс. чел. Была создана также система стандартов, некоторое количество общесистемных форматов и методических разработок. Кризис 1990-х гг. также почти полностью разрушил эту систему.

В настоящее время минимальными силами в Госкомстате и Госстандарте осуществляется поддержка (фактически только хранение) созданных когда-то классификаторов. Из новых общесистемных разработок следует отметить только появление в 1992-1994 гг. Общероссийского классификатора продукции и услуг, фактически перевод соответствующего классификатора ООН. Однако внедрение этого классификатора происходит крайне медленно. Можно ожидать, что сильным импульсом для модернизации и развития ЕСКК ТЭИ будет вступление России в ВТО, что потребует перевода многих российских систем на международные стандарты в области классификации.

ЛО автоматизированных библиотечных систем. В период бурного развития ЛО ГАСНТИ и ЕСКК ТЭИ в 1960-1980 гг. лингвистическое обеспечение библиотечных

систем находилось в зачаточном состоянии. Причина этому - слабое развитие АИС в библиотеках в те годы. Однако в 1990-е гг. библиотечное сообщество стало информатизироваться достаточно интенсивно и сейчас по уровню информатизации ничуть не уступает ГАСНТИ.

В то же время библиотечное сообщество не выдвинуло принципиально новых идей в отношении ЛО. Основные усилия направлялись на перевод в компьютерную форму уже принятых и адаптированных в библиотечном сообществе языковых средств, таких как ББК, УДК, библиографического описания или языки предметных рубрик.

Основным достижением следует считать принятие в качестве фактического стандарта русской версии MARC, а также определенные результаты по созданию “авторитетных файлов”, т.е. системы нормативных словарей для языков библиографических данных. Однако эти разработки предназначены только для библиотечного сообщества и по многим причинам не могут рассматриваться как универсальные решения для ЭБ.

ЛО архивных и музейных АИС. Информатизация архивов и музеев в значительной степени повторяет историю развития ГАСНТИ и библиотечных информационных систем, с отставанием на 10-20 лет. Поэтому попыток комплексного системного проектирования ЛО в архивах и музеях в советское время не было предпринято, а сейчас такой возможности, очевидно, нет. Тем не менее, в отдельных коллективах происходит довольно активное создание различных языковых средств, как аналогичных языкам, создаваемым в информационно-библиотечных системах, так и оригинальных.

ЛО систем искусственного интеллекта. Наиболее высокого уровня разработки ЛО в советское время были достигнуты в отдельных АИС, в той или иной степени использовавших идеи и методы искусственного интеллекта. Такие АИС создавались для более узких классов задач, чем АСНТИ, АБИС или обычные АСУ и все они носили уникальный характер. Описание системы подобного класса можно найти, например, в работах [9, 10].

Опыт создания систем класса искусственного интеллекта весьма ценен в теоретическом отношении, однако, трудно воспроизводим практически. Подобные разработки могли вестись только при сочетании видимых практических целей с серьезной академической базой, что вряд ли осуществимо в России в настоящее время.

ЛО негосударственного сектора. Параллельно с фактической ликвидацией общегосударственных систем ЛО ГАСНТИ и ЕСКК ТЭИ, в 1990-е гг. в России бурно развивались коммерческие и другие негосударственные информационные системы. Соответственно шло развитие и ЛО этих систем. В результате в отдельных компаниях были сделаны первоклассные разработки в области ЛО. Среди них следует отметить поисковые машины с применением морфологического анализа (Яндекс, Рамблер, Google, “Русский текст” и др.), системы навигации и поиска правовой информации (Гарант, Кодекс и др.), системы распознавания текстов и ведения машинных словарей (ABBYY), системы

распознавания устной речи (Cognitive Technology), системы машинного перевода (Промпт) и др. Однако заметных попыток интегрировать разработки в области ЛО, имея в виду достижения и прежних исследователей и современных систем в рамках крупных коммерческих АИС не было сделано.

Краткая характеристика современного состояния разработок ЛО. Наиболее продвинутыми являются в настоящее время средства ЛО коммерческих АИС. Однако по понятным причинам их разработки не являются ни системными, ни широко тиражируемыми.

В некоторых случаях эффективное создание ЛО происходит в настоящее время и в некоммерческих структурах. Так, наиболее полной и комплексной системой ЛО классического типа в настоящее время является ЛО УИС “Россия”, разработанная и реализованная в Московском государственном университете (руководитель – Т.Н.Юдина).

В целом можно констатировать отсутствие системных разработок в данной области. Реальная координация осуществляется только в библиотечном сообществе, однако достигнутый в нем уровень большинства разработок не позволяет претендовать на серьезное использование этих разработок за пределами библиотек.

Выдвинутая недавно претенциозная и масштабная идея программы “Электронная Россия”, требует для своей реализации адекватных усилий и в части разработки ЛО. Это вселяет надежду, что разработки ЛО в первое десятилетие нового века будут вестись более широким фронтом и с необходимым уровнем координации. Очевидно, что теория и практика создания ЛО ЭБ должна включать по возможности все достижения в этой области, имеющиеся как в государственном, так и в частном секторах, и обобщать положительный опыт и советского и российского этапов развития информатики.

Классификация средств ЛО. Исходя из изложенного, средства, входящие в состав ЛО, целесообразно разделить на 2 класса. К одному классу относятся языки, предназначенные непосредственно для представления данных в ЭБ. Говоря о данных, мы предполагаем, что они представлены в виде некоторых выделяемых и идентифицируемых информационных ресурсов, которые можно назвать “цифровыми объектами”.

Именно для этого класса языковых средств корректно применять широко распространенный термин *“информационно-поисковые языки”* (ИПЯ). Эти языки достаточно естественно классифицируются в зависимости от уровня отображения информации, имеющейся в цифровых объектах. Таких уровней можно выделить 4:

1. Уровень отображения цифрового объекта в целом, включая его формальные характеристики.

2. Уровень отображения тематики или содержания цифрового объекта

3. Уровень отображения семантики единиц естественного языка,

содержащихся в цифровом объекте.

4. Уровень отображения высказываний, содержащихся в цифровом объекте.

Для цифровых объектов типа документов первому уровню отображения соответствуют языки описания документов, весьма детально разработанные в традиционных областях информационной деятельности: библиотековедении, архивном деле, делопроизводстве, картографии и др.

Самый известный тип этих языков образуют *языки библиографических данных*, включающие правила библиографического описания и форматы библиографической записи. Эти языки возникли еще в XIX веке.

В настоящее время происходит активная интеграция языков библиографических данных с языками, применяющимися для описания других видов цифровых объектов. Особенно активно этот процесс развивается в Интернете. Общее название для языков, предназначенных для описания цифровых объектов – *системы метаданных.*

На втором уровне отображения используются языки *классификационного* или *предкоординатного* типа, также имеющие большую историческую традицию. Принципиальным свойством этих языков является разбиение множества цифровых объектов на классы, описанные при помощи априорного связывания (предкоординации) поисковых признаков этих классов, чаще всего, в виде иерархического дерева. Теория классификационных языков была предметом громадного числа исследований специалистов самого разного профиля. Можно даже вспомнить “классификационное общество”- неформальное сообщество ученых, интересовавшихся классификаций, активно функционировавшее в конце 1970 – в 1980-х гг.

Судьбы языков этого типа с учетом перспектив глобальных информационных сетей вызывают оживленные дискуссии, в связи с их имманентными недостатками, главный из которых – необходимость интеллектуального индексирования. При этом классификационные языки обладают заметными преимуществами перед другими типами поисковых языков, прежде всего наглядностью, простотой для пользователя и независимостью от естественного языка. В настоящее время классификационные языки являются обязательным компонентом практически всех крупных ЭБ.

Наиболее новым типом языковых средств, появившимся только в рамках автоматизированных систем в 1950-х гг. XX века, являются языки, ориентированные на использование в качестве лексики единиц естественного языка. Поэтому вполне адекватное название этой группы языков – *вербальные языки.* Однако наиболее распространенное название этих языков – *дескрипторные*, в соответствии с названием общепринятой формы представления лексических единиц этих языков (дескрипторов). Иногда эти языки также называют *посткоординатными*, подчеркивая противопоставление с классификационными языками по базовой функции – способу отражения информации текста. Если в классификационных языках используется априорное

связывание поисковых признаков, то в дескрипторных языках признаки связываются непосредственно в цифровом объекте (посткоординация).

Иногда можно встретить исследования, где вся проблематика лингвистического обеспечения сводится к проблематике вербальных языков, особенно при поиске по полным текстам документов. Это конечно слишком узкое представление, однако нет сомнения, что вербальные языки являются центральным компонентом лингвистического обеспечения ЭБ. Практически вся теория информационного поиска строится на использовании вербальных ИПЯ.

Большое развитие, по крайней мере, в АИС НТИ, получили языки, ориентированные на представление и поиск высказываний, точнее, фактов, содержащихся в документах. Этот класс языков находится на стыке АИС типа “электронной библиотеки” и АИС типа “банк данных”.

Поскольку основной и чуть ли не единственный тип высказываний, которые при разумных затратах удается автоматически извлекать из плохо структурированной информации – это факты типа “объект - признак - значение”, постольку языки данного класса принято именовать *“объектно-признаковыми”.* Иногда их также называют фактографическими или объектографическими. Следует иметь в виду, что такая терминология принята почти исключительно среди специалистов по электронным библиотекам, иначе говоря, специалистов по обработке плохо структурированной информации. В других направлениях информатики, прежде всего в теории баз данных, эти средства именуют “моделями данных”, языками описания данных, и др.

Существует класс еще более развитых, точнее семантически более сильных языков, переводящих обычные электронные библиотеки в класс систем искусственного интеллекта. В последние годы разработки языков этого класса заметно сократились. По состоянию на конец 1980-х гг. проблематика средств лингвистического обеспечения таких систем рассмотрена в работе В.Ш.Рубашкина [10].

Структура языков как системы знаков. Все перечисленные выше виды языковых средств можно с большей или меньшей степенью условности назвать языками. Однако, определив некоторый объект как язык, мы должны уметь выделять в его составе обязательные для любого языка компоненты. В любом языке выделяются знаковые единицы трех уровней:

- Алфавит – т.е. множество допустимых символов.
- Лексика – множество семантически интерпретированных знаков.
- Тексты (дискурс) – семантически интерпретированные знаковые единицы речи.

В любом языке также выделяются два класса правил (грамматики):

- Морфология – правила образования и изменения лексических единиц;
- Синтаксис – правила образования текстов.

Семантически интерпретированные знаковые единицы языка (лексика и тексты) обладают тремя типами отношений (свойств):

- Синтактика – отношения между знаками;
- Семантика – отношение знака к означаемому (денотату);
- Прагматика – отношение знака к участнику дискурса.

В теории и практике ЛО ЭБ эта схема обычно модифицируется. Алфавиты в большинстве случаев определяются программно-технологическими возможностями ЭБ и объектом проектирования в составе ЛО не являются.

Структура и особенности текстов на ИПЯ (поисковых образов документов и поисковых предписаний) обычно рассматривается как результат действий синтаксических правил, а не как самостоятельные знаки.

Под грамматикой ИПЯ обычно имеют в виду только синтаксис, морфологию ИПЯ, если она и выделяется, рассматривают на уровне лексики.

Сами синтаксические отношения обычно разделяются на два типа – синтагматические (отношения знаков в тексте) и парадигматические (отношения знаков вне контекста). Поскольку парадигматические отношения в реальных языках устанавливаются на уровне лексики, конкретно в словарях или классификациях, то эти отношения рассматриваются совместно с лексикой.

Таким образом, в составе ИПЯ реально выделяются два основных компонента – *лексика* (в том числе организованная в словари с использованием парадигматики) и *грамматика,* при помощи которой порождаются тексты на этих языках.

Что же касается прагматических свойств ИПЯ, связывающих тексты на ИПЯ с участником коммуникации, в данном случае поиска, то эти свойства реализуются в виде методик и алгоритмов индексирования, а также непосредственно в процессе поиска, при проектировании интерфейса, диалога пользователя с ЭБ, критериев ранжирования и выдачи результатов поиска.

Лингвистические процессоры. Второй класс средств, входящих в состав ЛО ЭБ, не является языками. Выше мы назвали их *лингвистическими процессорами*. Как мы уже отмечали, это достаточно широкий класс информационных технологий, но применительно к ЭБ к этим средствам целесообразно относить два класса технологий: системы автоматической обработки текста и лингвистические банки данных.

Под *автоматической обработкой текста* здесь понимаются процессы автоматического формирования описания текста (документа) на одном или нескольких информационных языках, включая автоматическое индексирование, аннотирование или реферирование. В основе этих процессоров лежат конкретные лингвистические алгоритмы, прежде всего морфологического и синтаксического анализа.

Лингвистические банки данных – важный обеспечивающий компонент развитых

ЛО ЭБ. Практически значительная доля затрат на создание и эксплуатацию ЛО – это затраты на создание и поддержание лингвистических банков данных. В этой части ЛО ЭБ смыкается с таким направлением информатики как компьютерная лексикография.

Итак, ЛО ЭБ включает следующие языковых средств:

1. Информационно-поисковые языки

1.1. Системы метаданных

1.2. Классификационные языки

1.3. Вербальные языки

1.4. Фактографические (объектно-признаковые) языки

2. Лингвистические процессоры

2.1. Системы автоматической обработки текста

2.2. Лингвистические банки данных

В заключение постараемся рассмотреть общее состояние и перспективы работ по лингвистическому обеспечению электронных библиотек в России. Применительно к разным типам языковых средств, рассмотренным в настоящей статье, эти перспективы видятся по-разному.

Центральной задачей для развития электронных библиотек в русскоязычном фрагменте Интернета видится, безусловно, развитие *систем метаданных*. Это направление наиболее интенсивно развивается в мировом Интернете и в значительной степени поддерживается крупнейшими производителями программных средств для Интернета, такими как Microsoft. Достаточно указать на усилия, затрачиваемые на создание и внедрение наиболее общей системы метаданных – Дублинского ядра. Системы метаданных определяют класс задач, которые реализуются в электронных библиотеках и решающим образом влияют на интероперабельность (совместимость) коллекций, имеющихся в ЭБ. Тем самым принятие тех или иных принципов в отношении метаданных фактически определит стоимость проектов по созданию электронных библиотек и эффективность затрат на эти проекты.

Важность задачи создания и внедрение эффективность систем метаданных определяют и фронт работ этого направления. В разной степени разработки и внедрения находятся десятки проектов – от глобальных и универсальных, типа Дублинского ядра, до более частных, например, создания диалектов XML для астрономии, архивов или финансовой отчетности.

В России это направление сильно отстает и в заметных масштабах работа практически не ведется. Отчасти это связано с тем, что проблема интероперабельности не очень затрагивает коммерческие электронные библиотеки и поисковые машины, задающие в настоящее время тон в разработках этого направления информатики. Государственной программы в этом направлении до сих пор не существует. Остается надеяться, что этой проблеме будет уделено должное внимание в рамках работ по Федеральной целевой программе “Электронная Россия”.

Что же касается конкретной концепции применения систем метаданных, то автор придерживается мнения, что единым и универсальным языком метаданных должен быть язык Дублинского ядра, который нужно принимать в качестве стандарта для всех электронных библиотек, создаваемых за счет бюджета.

Однако язык Дублинского ядра должен сосуществовать с другими, более развитыми языками, также основанными на XML, такими как ONIX, которые позволяют решать более частные задачи, например, для книготорговли. Таким образом, Дублинское ядро представляется как вершина иерархии систем метаданных, которая развивается более детально в конкретных коллекциях или сервисах системы ЭБ при помощи частных систем метаданных.

Что касается перспектив развития *классификационных систем*, то очевидно, что им предстоит значительная эволюция. Использование классификаций по многим причинам является обязательным. В то же время продолжающийся процесс вовлечения в Интернет и в электронные библиотеки разных классификационных систем, а также процесс создания новых классификаций пока существенно опережает обратный процесс унификации и конвергенции этих языков. Это означает продолжение роста разнообразия этих систем, хотя с точки зрения глобальной эффективности число различных классификаций, применяемых в Интернете, должно быть минимально.

Прогноз заключается в том, что пик роста разнообразия еще не наступил, хотя нет никакого сомнения, что рано или поздно этот перелом наступит. Мы предлагаем паллиативное решение, заключающееся в создании банка данных классификаторов, применяемых в российских электронных библиотеках. Создание такого банка данных позволит снизить величину разнообразия классификационных систем за счет объединяющей все классификации тезарусно-сетевой структуры, а также вспомогательных средств лексического поиска в банке данных классификаторов. В идеале при помощи такого средства возможен будет поиск по “своей” классификации в “чужом” массиве.

Вербальные языки были и остаются центральным элементом лингвистического обеспечения ЭБ. В настоящее время доминируют языки, основанные на свободной, неконтролируемой лексике и это вполне объяснимо. Однако уже многим разработчикам поисковых машин очевидны границы развития вербальных языков неконтролируемого типа. Сейчас эту проблему пытаются решить за счет параллельного использования классификационных поисковых языков типа традиционных каталогов.

Однако рано или поздно придется обратиться к идее семантически контролируемых поисковых языков, т.е. к идее тезауруса для Интернета или, по крайней мере, для контролируемой части информационного пространства Интернета, то есть для коллекций электронных библиотек. Специалистам в области ИПЯ такая перспектива очевидна уже давно, однако общая ситуация в области электронных библиотек пока не способствовала развитию “семантического” направления в ЛО ЭБ.

Следует добавить, что создание тезаурусов для поиска в Интернете уже начато за рубежом, хотя пока в академическом, а не в коммерческом секторе Интернета. Потенциал для такого развития событий имеется и в России. Однако для получения реальных результатов необходимо объединение усилий академических специалистов по информатике и коммерческих поисковых машин. В российских условиях это возможно только при специальных усилиях со стороны государства.

Серьезным подспорьем тезаурусному направлению ЛО ЭБ могло бы стать повышенное внимание к созданию общедоступных лингвистических банков данных в Интернете. Сейчас такие банки данных создаются либо для массового использования, либо для профессионалов-лингвистов. Необходимо поддержать специальные усилия по созданию лингвистических баз данных для нужд электронных библиотек.

Движение к “семантическим” ИПЯ послужит мощным стимулом для развития различных направлений в области *автоматической обработки текста*, которые сейчас ведутся чрезвычайно малыми и разрозненными силами. В то же время в этом направлении российский потенциал просто огромен. Не следует забывать, что в 1960-1980-х гг. российская прикладная лингвистика была одной из самых сильных, если не самой сильной в мире и достижения российских исследователей трудно преуменьшить. Даже просто повторить в современной программно-технологической среде результаты тех лет было бы большим достижением.

Однако решение этой проблемы, как и многих бед современной российской науки, лежит вне возможностей самих ученых. Без серьезного внимания к данному направлению со стороны лиц и организаций, способных организовать широкомасштабные работы, достижения российской прикладной лингвистики будут безвозвратно потеряны. Хотя широковещательные заявления первых лиц нашего государства об информатизации, как генеральной линии государственной политики в ближайшие годы внушают сдержанный оптимизм.

Литература

1. Армс В. Электронные библиотеки // М., ПИК ВИНИТИ, 2001

2. Антопольский А. Б, Вигурский К.В. Электронные библиотеки.// Информационные ресурсы России, 1999, № 4

3. Когаловский М.Р. Энциклопедия технологий баз данных. – М., Финансы и статистика, 2001

4. Михайлов А.И., Черный А.И., Гиляревский Р.С. Основы информатики - Наука, М., 1968

5. Белоногов Г.Г. Кузнецов Б.А. Языковые средства автоматизированных информационных систем—М.: Наука, 1983

6. Антопольский А.Б. Разработка и внедрение методов совместимости лингвистического обеспечения при взаимодействии АИС. // Дисс. на соиск. уч. степ. д.т.н. – М., 1990

7. Влэдуц Г. Э., Данилов М.П., .Уманский А.Н. Комплекс средств ндексирования научно-технической информации” (КСИНТИ) - ВНИИКИ, М., 1969

8. Положение о ЛО ГАСНТИ - ГКНТ- ВИНИТИ, М., 1986

9. Дракин В.И., Попов Э.В., Преображенский А.Б. Общение конечных пользователей с системами обработки данных.—М.: Радио и связь, 1988

10. Рубашкин В.Ш. Представление и анализ смысла в интеллектуальных информационных системах. - М., Наука, 1989

Об авторе

Антопольский Александр Борисович - доктор технических наук, чл.-корр. РАЕН, директор НТЦ "Информрегистр"

http://www.inforeg.org.ru/
