ТАТАРСКИЙ ЯЗЫК В КИБЕРПРОСТРАНСТВЕ

***Д.Ш. Сулейманов**, академик АН РТ, профессор*

В современном глобализирующемся мире, как никогда, остро встает вопрос сохранения и развития языков. Причем не только как когнитивного механизма, т.е. как средства концептуализации и описания воспринимаемой человеком действительности, и не только как коммуникативного механизма, т.е. средства общения и достижения взаимопонимания, но, что не менее важно, и как механизма передачи знаний из поколения в поколение, как средство передачи опыта народа, накопленного тысячелетиями и доступного только на языке этого народа. Вместе с тем, ситуация с языками иначе как тревожной не назовешь. По некоторым оценкам, половина из ныне существующих порядка 6000 языков прекратит существование к 2050 г. Как отмечает заместитель Генерального директора ЮНЕСКО по коммуникации и информации А.В. Хан в предисловии к книге [1], из всего многообразия существующих языков только 12 используются для написания 98% веб-страниц. Причем, судя по анализу специалистов, представители даже таких мощных, как славянские, тюркские, финно-угорские языки не могут быть спокойны за свое будущее. По оценке О'Неля, Лавуа и Бене, в 2003 г. 72% всей информации в Интернете размещалось на английском языке, и это доминирование в глобальной сети только с годами увеличивается. И даже те нации, языки которых обладают некоторыми признаками активности, являются ежедневным средством самовыражения и общения довольно большого круга населения, используются в средствах массовой информации, имеют формальные признаки поддержки со стороны государства (государственный язык, отдельная Госпрограмма поддержки татарского языка в различных сферах), не могут быть уверены в своем будущем, если языки активно и полноценно не включены в систему образования и общественную сферу в целом. Сегодня в киберпространстве используются менее ста языков [1].

Очевидно, существование языка, а также его развитие и дальнейшее сохранение как культурного явления в значительной степени зависят от активности языка в компьютерных информационных технологиях.

В статье данный тезис рассматривается на примере внедрения татарского языка в так называемое киберпространство (то есть пространство взаимодействия человека и компьютерных систем и технологий). При этом, как будет показано ниже, обеспечение функционирования татарского языка в компьютерных системах и технологиях является актуальным не только в плане повышения его активности и конкурентоспособности среди других языков в качестве средства накопления информации и общения с компьютером, но и в плане создания новых технологий хранения и обработки информации на основе татарского языка в силу целого ряда когнитивных особенностей его структуры и лексического корпуса.

Очевидно также, что для обеспечения равного функционирования татарского и русского языков как государственных в Республике Татарстан, необходимо, чтобы татарский язык так же, как и русский, стал рабочим языком компьютеров. Соответственно, наряду с задачей использования татарского языка в инфокоммуникационных технологиях и создания специальных программ

обработки татарского языка, ставится также задача татарской локализации их интерфейсной оболочки, т.е. средств общения компьютера с человеком.

Исследования и разработки по внедрению татарского языка в компьютерные технологии в Республике Татарстан начались практически с конца 1980-х годов, с разработки первых драйверов периферийных устройств, текстового редактора и татарского корректора, необходимых для компьютерного издания татарских книг, газет и журналов и ведения делопроизводства. В 1993 г. для решения задач в рамках научно-прикладной программы Академии наук РТ «Компьютерное обеспечение функционирования татарского языка как государственного» и для разработки средств компьютерного обеспечения татарского языка как государственного в рамках Государственной программы РТ по сохранению, изучению и развитию языков народов Республики Татарстан была создана Совместная научно-исследовательская лаборатория Академии наук РТ и Казанского государственного университета «Проблемы искусственного интеллекта». В настоящее время в соответствии с Постановлением Кабинета Министров РТ № 590 от 28.08.2009 на основе СНИЛ ПИИ АНТ и КФУ создан НИИ «Прикладная семиотика», работающий в составе Отделения социально-экономических наук АН РТ. В рамках института выполняются следующие фундаментальные и прикладные задачи: разработка лингвистического и программного обеспечения интеллектуальных информационных систем (машинный перевод для родственных языков, онтолингвистическое моделирование, обеспечение внедрения татарского языка в современные инфо-коммуникационные технологии, речевые технологии); разработка концептуальных моделей научных и образовательных электронных ресурсов (электронный корпус татарского языка, виртуальный музей-библиотека членов АН РТ); семиотическая база данных в гуманитарных приложениях, мультимедийные электронные учебники; когнитивные исследования инфо-коммуникационной образовательной среды (интеллектуальная система управления контрольно-диагностическим компонентом в образовательной среде, интеллектуальный потенциал общества, организации и личности, педагогика электронного образования, электронные учебники нового поколения).

Фундаментальные исследования и прикладные разработки по поддержке татарского языка в информационных технологиях изначально осуществляются в трех основных направлениях:

1) внедрение татарского языка в киберпространство,

2) разработка и адаптация информационных технологий для татарского языка,

3) использование когнитивных возможностей татарского языка для создания новых информационных технологий.

1. Внедрение татарского языка в киберпространство

Первое направление исследований и разработок непосредственно связано с проблемой сохранения языка, повышения его активности в мировом инфокоммуникационном пространстве, использования татарского языка в киберпространстве как когнитивного и как коммуникативного средства, т.е. средства представления, накопления и передачи информации, обеспечения паритетного функционирования татарского и русского языков как государственных в Республике Татарстан, а также предоставления возможности носителям языка прямого общения с компьютерными системами без языка-посредника. Данное направление работ включает базовую и полную локализации компьютерных систем, то есть адаптацию их под татарский язык.

В настоящее время эта задача решена в полном объеме для татарского языка на основе кириллической графики. Учеными Академии наук РТ и КФУ разработаны экранные и клавиатурные драйверы, драйверы печати и шрифто-

вое обеспечение для татарского языка на кириллической основе и предложены в качестве стандарта для применения в информационных технологиях в Республике Татарстан. На их основе принято Постановление КМ РТ «О стандартах кодировки символов татарского алфавита для компьютерных применений» (№ 1026 от 9 декабря 1996 г.).

Данное Постановление помогло унифицировать драйверы устройств, которые в первое время создавались различными группами и отдельными специалистами по своему усмотрению и практически, как вирус, распространились по различным компьютерам, закрепляя разную раскладку одних и тех же татарских букв на кодовых страницах, создавая «разночтение». Унификация кодовой страницы и драйверов устройств, помогла ликвидировать начавшийся хаос в делопроизводстве, когда татарские тексты, набранные на одной машине, на другой – не читались или отображались некорректно.

На базе принятых стандартов по соглашению с фирмой MicroSoft были разработаны соответствующие драйверы устройств и внедрены в операционную среду Windows NT и Office-2000. В настоящее время пакет драйверов TATWIN, включенный в программный комплекс поддержки татарского языка TatSoft 2, позволяет вести делопроизводство на татарском языке на кириллической основе во всех приложениях операционной системы WINDOWS'95, '98, '2000, 'XP, Vista, 7, а также работать в Интернете. Соответствующая информация имеется на web-сайте фирмы Майкрософт. Таким образом, татарский язык стал вторым тюркским языком (после турецкого языка), подготовленным для реализации специалистами самой республики (а не разработчиками фирмы) и доступным в среде Windows при ее инсталляции на любом рабочем месте.

Сотрудничество Академии наук РТ с Московским бюро фирмы Майкрософт, начавшееся уже в 1995 г. с татарской локализации ОС Windows'95, нашло перспективное продолжение. В 2005–2010 гг. осуществлена полная татарская локализация основных продуктов фирмы Майкрософт Windows XP, Windows Vista, Windows 7 и офисных приложений. Научно-исследовательским институтом «Прикладная семиотика» Академии наук РТ и лабораторией «Проблемы ИИ» КФУ разработан татарский интерфейс операционной системы и, таким образом, татарский язык, наряду с такими мировыми языками, как английский и русский, стал родным языком для операционной системы Windows и таких активно используемых пользовательских программ, как Word, Exel, Power Point.

Татарская локализация операционной среды MS Windows и ее приложений ведет к активному внедрению татарского языка в инфокоммуникационные технологии, развитию татарского языка и распространению его в мировом информационном пространстве. Очевидно, что, только становясь языком компьютерных технологий, языком накопления, обработки, передачи информации, языком общения с компьютерными системами, татарский язык, впрочем, как и языки других народов, имеет возможность стать полнокровным государственным языком в республике, языком культуры, языком науки, языком общения в киберпространстве.

Разработана также опытно-эксплуатационная версия пакета драйверов и шрифтового обеспечения для татарского языка на основе латиницы. По представлению Академии наук РТ, принято Постановление КМ РТ «О стандартах кодировки символов татарского алфавита на основе латинской графики и базовых программах для компьютерных применений» (№ 625 от 27 сентября 2000 года). Разработан программный пакет TATLAT, позволяющий вести делопроизводство, издательское дело на татарском языке на основе латинской графики во всех приложениях операционной системы Windows'95 '98 '2000 'NT4. Однако активное внедрение этих

пакетов программ, продолжавшееся в течение нескольких лет после принятия соответствующих документов закона РТ, указа Президента РТ, постановления Кабинета Министров РТ, Государственной программы реализации, а также множества образовательных инициатив, в настоящее время приостановлено по известному решению Госдумы РФ.

2. Разработка и адаптация информационных технологий для татарского языка

В рамках *второго направления* разработаны пакеты прикладных программ для работы с татарским языком, программные средства, позволяющие компьютеризировать делопроизводство, издание газет и журналов, проверять корректность татарских текстов, автоматизировать рабочие места специалистов. Осуществляются исследования теоретических и прикладных проблем компьютерной лингвистики применительно к татарскому языку, к его грамматике, лексикологии и лексикографии, к различным проявлениям в речи, с целью построения прагматически-ориентированных лингвистических моделей и создания на их базе систем автоматизированной обработки татарского языка. Важными и активно разрабатываемыми и, очевидно, судьбоносными являются вопросы татарской терминологии в киберпространстве.

В настоящее время создана полнофункциональная компьютерная модель морфологии татарского языка, причем, учитывая структурную специфику татарского языка и исходя из прикладных задач, разработаны три различные модели морфологии. Генеративная модель морфологии, основанная на правилах словоизменения, хотя и уступает другим моделям по быстродействию, обеспечивает полноту анализа словоформы, позволяя в полной мере учитывать агглютинативный характер языка, распознавая словоформы потенциально неограниченной длины. Парадигматическая модель татарской морфологии обеспечивает быстрое распознавание словоформ и анализ корректности татарских словоформ с точностью до 95%, используется в поисковой системе УИС «Россия» (ЦИТ МГУ, г. Москва) и в среде MS Windows и ее офисных приложениях. Причем скорость распознавания составляет 100 слов в 0.014 секунд, что перекрывает требования заказчика на целый порядок. Кроме того, в рамках совместного проекта с Белкентским университетом (Турция) разработана двухуровневая модель морфологии татарского языка, реализованная в среде известной программной оболочки PC KIMMO и используемая в составе татарско-турецкого машинного переводчика. Создана также структурно-функциональная модель татарских аффиксальных морфем, являющаяся «инвентарной базой» для построения различных прагматически-ориентированных морфологических моделей и на ее базе построен интегрированный программно-информационный комплекс «Татарская морфема». Данный комплекс практически является автоматизированным рабочим местом (АРМом) для разработчиков различных лингвопроцессоров, а также для осуществления учебно-исследовательской деятельности в татарском языкознании, может быть успешно использован как исследовательский инструмент и в других языках.

Еще одна полезная программа – татарско-русский машинный переводчик татарских фамильно-именных групп, созданная на основе словаря компонент и правил, учитывающих специфику образования татарских собственных имен, является незаменимым инструментом в автоматизированных системах ЗАГС и Паспортно-визовой службы, а также для автоматической генерации татарских имен и фамилий на основе модели компонент татарского имени. Специалистами института осуществлена татарская локализация оптического распознавателя текстов FineReader московской фирмы ABBYY. Данная программа благодаря встроенной морфологии татарского языка

распознает татарские тексты с такой же точностью и быстротой, как русские и английские.

Важной задачей, которая выполняется институтом, является создание и поддержка электронного корпуса татарского языка, практически представляющего собой машинный фонд татарского языка (МФТЯ) в сети Интернет со следующими корпусами: а) электронные неформатированные тексты (газеты, журналы, книги, документы и др.); б) размеченные тексты, словари, тезаурусы; в) программные модули: лингвопроцессоры (машинные переводчики, синтезатор речи, распознаватель текста и речи и др.), АРМы специалиста (учителя, редактора, лингвиста и др.), интеллектуальная многоязычная машина поиска. Задача создания электронного корпуса татарского языка является фундаментальной научно-практической проблемой, решение которой даст возможность быстрого и удобного доступа к различным лингвистическим ресурсам большого объема посредством использования вычислительных машин. Реализация данного проекта приведет к формированию соответствующей инфраструктуры (татарский контент и средства работы с татарским контентом) для полноценного представления татарского языка в сети Интернет.

Одним из интересных и полезных продуктов, разработанных институтом совместно с фирмой ABBYY и ИЯЛИ АН РТ, является Многоязычный электронный словарь Lingvo'x3 с татарским языком, представляющий собой практически настольную библиотеку из 154 различных словарей на 12 языках мира, в числе которых имеется и татарский язык. Ценность данного электронного словаря для татароязычного пользователя, кроме многих других возможностей, заключается в том, что через татарско-русскую языковую пару доступны переводы во всех 154 словарях на 11 языках мира (то есть, включив татарско-русский словарь объемом порядка 60 000 словарных статей, потенциально мы получили татарско-английский, татарско-французский, татарско-испанский, татарско-немецкий, татарско-китайский, татарско-турецкий и др. двуязычные словари).

Незаменимым инструментом в делопроизводстве и издательском деле является программа WordCorr – морфологический корректор татарских текстов для Microsoft Word, который позволяет находить и исправлять ошибки в татарских текстах, при этом предлагая возможные корректные варианты. Функционирует во всех операционных системах MS Windows'95 '98 '2000 'XP, Vista, 7 и приложениях, причем в последнем продукте татарский спелчеккер является встроенным и разработан совместно со специалистами фирмы Майкрософт.

Практически с 1990-х годов осуществляется активная работа по разработке электронных обучающих программ татарскому языку, а также программ обучения предметов на татарском языке. Ряд последних разработок доступны в Интернете, среди них: Татар Телле Заман (ТТЗ) – мультимедийный электронный учебник по татарскому языку (в процессе заполнения контента и доработки некоторых блоков контроля и лингвистических игр) (http://ttz.fosslabs.ru/), Татар-онлайн – мультимедийный Интернет-учебник по татарскому языку (http://dev.tol.tatar.ru/), мультимедийный учебник 5 класса для дистанционного Интернет-обучения татарскому языку (http://distat.stage.metastudio.ru/).

Локальный вариант программы «Татар Телле Заман» уже реализован в виде компакт-диска и активно используется в школах республики. Программа содержит более 2000 татарских слов, более 2500 рисунков и фотографий, озвученные диалоги на различные темы и 11 увлекательных лингвистических игр, три типа различных упражнений, позволяющих тестировать знания обучаемого, возможности для совершенствования татарского произношения вслед за диктором. Многоязычный интерфейс (русский, татарский *(кириллица, латиница)*,

английский) системы позволяет изучать татарский язык как в русскоязычной, так и англоязычной среде. Татарские версии электронных мультимедийных учебных пособий Химия-8 и Физика-7, разработанные совместно с московской фирмой «Просвещение-Медиа» при содействии Министерства образования и науки РТ и Издательства «Магариф», благодаря комплексу разнообразных мультимедийных возможностей (видеосюжеты, анимация, звук, качественные иллюстрации, сотни интерактивных заданий и т.д.) обеспечивают увлекательный и эффективный процесс обучения. Разработано и передано в школы республики электронное мультимедийное учебно-методическое пособие «Татар теле-5». Электронное пособие содержит учебный материал по 6 темам, 123 упражнения, разделенных на 27 типов; включает гипертекстовый справочный материал по татарскому языкознанию, руководство пользователя и анимационную контекст-подсказку по запросу пользователя в он-лайн режиме. Программное обеспечение и технологии разработки и реализации мультимедийных учебных пособий, разработанные с ориентацией на татароязычную среду, в основе своей являются универсальными, независимыми от языка и успешно могут быть использованы также при создании электронных учебных пособий для других проблемных областей и для других языков.

В настоящее время нами реализован пакет программ поддержки татарского языка в инфокоммуникационных технологиях TatSoft 2, который включает перечисленные выше программные средства работы с татарским языком и дает возможность любому пользователю инсталлировать их на своем компьютере.

Среди перспективных работ института в плане поддержки татарского языка в компьютерных технологиях можно выделить следующие проекты.

1. Разработка Интеллектуальной многоязычной поисковой машины (ИМПМ). Актуальность работ по созданию ИМПМ связана с необходимостью создания машинного фонда (ресурса электронных коллекций) татарского языка, сложившейся языковой ситуацией в Республике Татарстан, появлением новых лингвистических и интеллектуальных технологий многоязыкового поиска, основанных на глубоком разрешении лексической многозначности. Кроме того, потребность в многоязыковых поисковых технологиях обусловлена тем фактом, что ряд развитых государств имеют несколько официальных языков, что дает проекту перспективу дальнейшего коммерческого использования.

2. Разработка программы распознавания татарской речи (включая русскую и английскую). Как прогнозируется специалистами, одним из основных направлений развития в сфере высоких технологий в ближайшие годы будут речевые технологии, особенно, автоматическое распознавание речи (АРР). Ожидается широкое внедрение технологий АРР в ведущие сектора экономики. По оценкам аналитиков, объем рынка продукции, использующей АРР, будет сравним с рынками таких высокотехнологичных товаров, как микропроцессоры, персональные компьютеры, программное обеспечение.

3. Разработка татарско-русского машинного переводчика, а также машинных тюркоязычных переводчиков в паре с татарским языком. Если особая актуальность машинных переводчиков первой группы объясняется необходимостью доступа к англоязычным базам знаний в Интернете через русский язык (априори предполагается, что русско-английский переводчик имеется) и необходимостью поддерживать равное функционирование татарского и русского языков как государственных в Республике Татарстан, то вторая группа – среди родственных языков – привлекательна в силу относительной простоты и малой затратности решения этой задачи (в некоторых случаях практически это простая конвертация текстов, например, для татарско-башкирской

пары языков), а также благодаря культурологической функции такого переводчика, помогающего сближению родственных народов.

3. Использование когнитивных возможностей татарского языка для создания новых информационных технологий

Третье направление исследований связано с актуальной задачей создания интеллектуальных операционных систем и интеллектуального программного инструментария на основе использования потенциала естественных языков, их семантических и синтаксических конструкций, а также лексического корпуса. Очевидно, что естественный язык является основой любой символической системы, определенным образом организованной, имеющей свой синтаксис и свою семантику (сюда же включается любая логика, математика и др.). Соответственно вместе с языком в этих системах реализуется и ментальность языка (точнее, ментальность этноса, передаваемая через язык).

Что является важным для компьютерных технологий? Известно, что критичными, соответственно важными для компьютерных технологий являются такие показатели, как *время обработки* информации, *объем памяти* для хранения информации (сжатие информации), *активность знаний* и возможность *задания нечетких команд* (однозначно воспринимаемых в определенном контексте). Последние два свойства являются необходимыми характеристиками для интеллектуальных систем и технологий. В связи с этим весьма актуальными и перспективными являются когнитивные исследования в языке с целью определения таких структур, схем, формул, которые в естественном языке реализуют указанные свойства и могут быть эффективно использованы при создании искусственных языков и систем программирования, а также любых других средств описания, хранения и обработки информации.

Как известно, операционные системы, языки программирования, средства обработки информации, практически все программное обеспечение, используемое сегодня, разработаны на основе английского языка и соответственно на основе менталитета английского языка (менталитета, отражаемого через английский язык – западного менталитета). Английский язык является языком флективно-аналитического типа (флексия – когда допускается и префиксное, и инфиксное и постфиксное изменение формы слова; аналитический тип – когда новое значение образуется сочетанием слов), практически с нулевой морфологией (по сравнению с агглютинативными языками). Отсюда следует, что сложный смысл образуется словосочетаниями и это приводит к большой комбинаторике при анализе. А это, в свою очередь, ведет к увеличению самых критичных показателей в вычислительных системах – объема требуемой памяти и времени при обработке информации. Выход из такой ситуации – исключение большого контекста, глубины конструкций, в итоге – упрощение смысла, семантики, соответственно и «интеллектуальных показателей». Таким образом, в основе самого английского языка заложен «интеллектуальный» тупик для вычислительных машин, заставляющий их не умнеть, а искать выход через повышение быстродействия системы и увеличение памяти, т.е. через развитие «физики», а не «мозгов».

Еще один недостаток технологий, основанных на английском языке, заключается в том, что сам строй языка, его синтаксис, «сопротивляется», даже противоречит одному из главных свойств интеллектуальности системы – *активности знаний*. Английский язык относится к языкам типа SVO (Subject-Verb-Object). То **есть «Субъект: Действие – Информация» (***I'll go to the cinema tomorrow afternoon with my friend ...*). Таким образом, сначала требуется выполнить, потом рассуждать, анализировать. Решение принимается не на основе информации, а информация подается в рамках выбранного действия, то есть не информация диктует, какое именно

действие необходимо совершить, какие методы, алгоритмы применять для ее обработки, а наоборот, действие, средство, схема, алгоритмы заставляют форматировать, структурировать, модифицировать информацию.

В отличие от индо-европейских языков, тюркские языки относятся к языкам типа SOV (Subject – Object–Verb). Соответственно реализуется схема: «Субъект: Информация – Действие». Например, смысл английского предложения, приведенного выше, будет передаваться следующим татарским предложением: *Мин ... иртәгә төштән соң дустым белән кинога барам*. То есть в татарском предложении сначала раскрывается информация, анализ ситуации, а затем уже в конце предложения приводится действие, отображаемое, как правило, глаголом.

Как показывают исследования, проводимые в НИИ «Прикладная семиотика» АН РТ, тюркские языки как агглютинативные языки, обладающие регулярной морфологией [2] и вместе с тем естественной сложностью, разрешаемой по контексту, являются эффективным инструментом для создания интеллектуальных систем обработки информации [3–5]. В силу минимальных показателей временных и емкостных оценочных функций для генерации и анализа цепочек словоформ (за счет регулярности) достигается оптимальность при накоплении и обработке информации. Компактность передачи смысла текста на поверхностном, лексическом, уровне объясняется также возможностями языка синтетически, т.е. словоформой, кодировать смысл, который для других языков (английский, русский) формируется аналитически, чаще всего, несколькими предложениями.

Агглютинативность языка, алгоритмические закономерности, минимальность исключений, наличие мощного мета-аппарата, достаточная жесткость синтаксиса и активность информации позволяют ставить задачу возможности построения языка промежуточной трансляции, т.е. языка-посредника на базе татарского языка, и даже разработки новых операционных систем на основе новой идеологии.

ПРИМЕЧАНИЯ

1. ЮНЕСКО. «Программа для всех». Как обеспечить присутствие языка в киберпространстве? –М.: Межрегиональный центр библиотечного сотрудничества (МЦБС) 2007. – 64 с.

2. *Heintz J.* and *Schonig C.* Turcic Morphology as Regular Language // Central Asianic Jornal (CFJ), 1989. – P.1–24.

3. *Suleymanov D.S.* Natural possibilities of the Tatar morphology as a formal base of the NLP // In Proceedings of the First International Workshop «Computerisation of Natural Languages» (Varna, Sept. 3-7, 1999). –Sofia (Bulgaria): Information Services Plc, 1999. – P.113.

4. *Сулейманов Д.Ш.* Естественные когнитивные механизмы в татарском языке //Тр. Междунар.семинара Диалог-2002 «Компьютерная лингвистика и интеллектуальные технологии» (г.Протвино, 6–11 июня 2002 г.): в 2 т. / Под ред. А.С.Нариньяни. – М.: Наука, 2002. – С. 500–507.

5. *Suleymanov D.S.* Natural cognitive mechanisms in the Tatar language // In the Collection of the Vienna Proceedings of the Twentieth European Meeting in Cybernetics and Systems Research. Edited by Robert Trappel. Vienna, Austria, 6–9 April, 2010. – P. 210–213.