– IT В ИССЛЕДОВАНИИ АЛТАЙСКИХ ЯЗЫКОВ –

DOI: 10.25587/2782-6627-2023-3-60-69 УДК 004.8+81'322

Современные лингвистические ресурсы и IT-разработки для татарского языка: региональный опыт

Д. Ш. Сулейманов, Р. А. Гильмуллин, А. Р. Гатиатуллин
Академия наук Республики Татарстан, Институт прикладной семиотики, г. Казань, Россия
E-mail: dvdt.slt@gmail.com
E-mail: rinatgilullin@gmail.com
E-mail: ayrat.gatiatullin@gmail.ru

Аннотация. Вопрос сохранения языка этноса как языка коммуникации, а также как носителя культурных кодов в настоящее время обрел особую остроту и актуальность. Как сохранить и развивать язык в условиях глобализации и повсеместной цифровизации, в условиях наступающего, а во многом уже и реального, информационного хаоса и в условиях, когда в ближайшей перспективе большинству из известных языков грозит полное исчезновение. Очевидно, что одним из надежных и перспективных решений является включение языка в цифровое пространство как языка накопления, обработки и передачи информации; языка коммуникации в глобальной компьютерной сети Интернет. В настоящее время в Республике Татарстан практически полностью реализован комплекс организационных и директивных мер и создана необходимая информационно-технологическая база для обеспечения полноценного функционирования татарского языка в компьютерных технологиях, включая деловую сферу и сферу управления, издательское дело, образовательную деятельность, лингвистические научные исследования и прикладные разработки. Еще одной действенной и весьма перспективной активностью являются исследование и использование языка как когнитивно-концептуального инструментария, способного обогатить ментальное и духовное пространство цивилизации в качестве лингво-семиотического ресурса, имеющего необходимый потенциал для создания новых технологий обработки знаний и интеллектуального интерфейса. В данной работе представлен опыт научно-исследовательской и практической деятельности Института прикладной семиотики Академии наук Республики Татарстан в одном из важных направлений искусственного интеллекта,

---

СУЛЕЙМАНОВ Джавдет Шевкетович – д. т. н., профессор, академик АН РТ, г. н. с. Академии наук Республики Татарстан, Институт прикладной семиотики.

E-mail: dvdt.slt@gmail.com

SULEYMANOV Djavdet – Doctor of Technical Sciences, Professor, Chief Scientific Associate, Academy of Sciences of the Republic of Tatarstan, Institute of Applied Semiotics.

ГИЛЬМУЛЛИН Ринат Абрекович – к. ф.-м. н., директор, Академия наук Республики Татарстан, Институт прикладной семиотики.

E-mail: rinatgilullin@gmail.com

GILMULLIN Rinat – Candidate of Physical and Mathematical Sciences, Director, Academy of Sciences of the Republic of Tatarstan, Institute of Applied Semiotics.

ГАТИАТУЛЛИН Айрат Рафизович – к. т. н., в. н. с., Академия наук Республики Татарстан, Институт прикладной семиотики.

E-mail: ayrat.gatiatullin@gmail.ru

GATIATULLIN Ayrat – Candidate of Technical Sciences, leading researcher, Academy of Sciences of the Republic of Tatarstan, Institute of Applied Semiotics.

именуемого «Обработка естественного языка» (NLP). Исследования и разработки осуществляются в трех научно-прикладных направлениях: 1) татарская национальная локализация инфокоммуникационных технологий; 2) разработка программного инструментария, систем и технологий обработки естественного языка (ЕЯ) и создание лингвистических ресурсов; 3) создание новых интеллектуальных технологий обработки информации на основе исследования когнитивного потенциала татарского языка (ТЯ).

Ключевые слова: инфокоммуникационные технологии, татарский язык, терминология, лингвистические ресурсы, когнитивный потенциал языка.

Для цитирования: *Сулейманов Д. Ш., Гильмуллин Р. А., Гатиатуллин А. Р.* Современные лингвистические ресурсы и IT-разработки для татарского языка: региональный опыт // DOI: 10.25587/2782-6627-2023-3(10)-60-69.

Modern linguistic resources and IT applications for the Tatar language: Regional practices

D. Sh. Suleymanov, R. A. Gilmullin, A. R. Gatiatullin
Academy of Sciences of the Republic of Tatarstan, Institute of Applied Semiotics, Kazan, Russia
E-mail: dvdt.slt@gmail.com
E-mail: rinatgilullin@gmail.com
E-mail: ayrat.gatiatullin@gmail.ru

Abstract. The issue of preserving the language of an ethnic group as a language of communication, as well as a carrier of cultural codes, has now acquired particular urgency and relevance. How to preserve and develop a language in the conditions of globalization and widespread digitalization, in the conditions of coming, and in many ways already real, information chaos and in conditions when in the near future most of the known languages are threatened with complete extinction. Obviously, one of the reliable and promising solutions is embedding a language in the digital space as a language of accumulation, processing and transmission of information; a language of communication in the global Internet network. Currently, in the Republic of Tatarstan, a set of organizational and policy measures has been almost completely implemented and the necessary information and technological base has been created to ensure full functioning of the Tatar language in computer technologies, including business and management spheres, publishing, educational activities, linguistic research and development. Another effective and very promising activity is the research and use of the language as a cognitive-conceptual tool that can enrich the mental and spiritual space of civilization as a linguistic-semiotic resource that has the necessary potential for creating new knowledge processing technologies and an intelligent interface. This paper presents the experience of research and practical activities of the Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan in one of the important areas of artificial intelligence called “Natural Language Processing”. Research and development are carried out in three theoretical and applied areas: 1) Tatar national localization of infocommunication technologies, 2) Development of software tools, systems and technologies for natural language processing and creation of linguistic resources, 3) Creation of new intelligent information processing technologies based on the research the cognitive potential of the Tatar language.

Keywords: infocommunication technologies, Tatar language, terminology, linguistic resources, cognitive potential of the language

For citation: *Suleymanov D. Sh., Gilmullin R. A., Gatiatullin A. R.* On the experience of implementing the Tatar language into infocommunication technologies in the Republic of Tatarstan // DOI: 10.25587/2782-6627-2023-3(10)-60-69.

Введение

Исследования и разработки Института прикладной семиотики Академии наук Республики Татарстан в одном из важных направлений искусственного интеллекта, именуемого «Обработка естественного языка» (NLP), осуществляются в трех научно-прикладных направлениях: 1) татарская национальная локализация инфокоммуникационных технологий; 2) разработка программного инструментария, систем и технологий обработки естественного языка (ЕЯ) и создание лингвистических ресурсов; 3) создание новых интеллектуаль-

ных технологий обработки информации на основе исследования когнитивного потенциала татарского языка (ТЯ). В статье описывается опыт исследований и разработок Института прикладной семиотики Академии наук Республики Татарстан в области сохранения, изучения и развития татарского языка с использованием инфокоммуникационных технологий по данным трем направлениям.

Первое направление ориентировано на решение задач сохранения татарского языка, повышения его активности в сети Интернет, а также использования ИТ как когнитивного и коммуникативного средства. Наиболее наукоемким, требующим тщательной практической проработки, являются создание стандартов для использования татарского языка в компьютерных системах, создание и стандартизация системы татарских терминов и понятий в информационных технологиях.

В рамках второго направления создаются технологии и системы, программный инструментарий и лингвистические ресурсы, предназначенные, главным образом, для обработки тюркских языков. В статье приведено описание ряда разработок, наиболее наукоемких и нетривиальных в функциональном и технологическом планах: интернет-платформа «Многофункциональная модель тюркской морфемы (МТМ)», представляющая собой инструментальную среду и базу тюркских аффиксальных морфем для создания лингвистических ресурсов и систем обработки тюркских языков; русско-татарский машинный переводчик и система многоязычного машинного перевода для 7 тюркских языков; системы синтеза и распознавания татарской речи; татарский национальный корпус «Туган тел» (tugantel.tatar) – лингвистический ресурс современного литературного татарского языка, представляющий собой уникальную инструментальную платформу, разработанную специалистами Института.

Третье направление – фундаментальные исследования когнитивного потенциала татарского языка для создания интеллектуальных технологий Ключевая идея – разработка и использование семиотических моделей лексико-грамматических конструкций тюркских языков как формальной основы интеллектуальных технологий, а также лексического корпуса как базы данных. Среди важных признаков интеллектуальности систем принятия решений, как правило, выделяются такие свойства, как активность знаний, то есть первичность анализа данных и вторичность принятия решения на основе этого анализа; возможность оперировать нечеткой информацией, семантически управляемой контекстом, и исполнять нечеткие команды. На конкретных примерах показывается перспектива исследования в этих целях татарского языка как языка агглютинативного типа, обладающего такими важными свойствами, как регулярность, рекурсия, активность знаний и ряд других. Работы этого направления создают перспективу построения универсального языка общения систем ИИ между собой и с человеком, а также интеллектуальной операционной системы, как когнитивно-коммуникативной системы ИИ.

Татарская локализация ИКТ

Направление исследования и разработки в этой области ориентированы на решение проблем сохранения языка, повышения его активности в сети интернет; использования татарского языка в ИТ как когнитивного и коммуникативного средства; обеспечения функционирования татарского языка как государственного в Республике Татарстан, а также национальной локализации интерфейса и ресурсов ИТ. Наиболее наукоемкими, требующими фундаментальных исследований, а также тщательной практической проработки, являются задачи разработки системы татарских терминов и понятий, создания стандартов для использования татарского языка в информатике и информационных технологиях [1].

В настоящее время эта задача решена в полном объеме для татарского языка на основе кириллической графики. По предложению Академии наук РТ принято Постановление

КМ РТ «О стандартах кодировки символов татарского алфавита для компьютерных применений» № 1026 от 9 декабря 1996 года. На его основе разработаны экранные и клавиатурные драйверы, драйверы печати и шрифтовое обеспечение для татарского языка на кириллической основе. Создан пакет стандартных технологий и программ базовой татарской локализации «Татарский Офис 2001». Унификация кодовой страницы и, соответственно, драйверов устройств, помогла ликвидировать начавшийся хаос в текстовых редакторах и издательских системах, когда татарские тексты, набранные на одной машине, не читались на другой или отображались некорректно.

На базе принятых стандартов по соглашению с фирмой Microsoft осуществлена полная татарская локализация операционной системы MS Windows и ее офисных приложений, включая интерфейс и справочные файлы, а также корректор татарских текстов (совместно с фирмой Microsoft). Разработано и внедрено более 5000 новых татарских терминов и понятий для компьютерных технологий. Текст переводов на татарский язык интерфейса и файлов помощи составил порядка 1 млн словоформ. Работа по татарской локализации новых версий программных продуктов Microsoft продолжается и сегодня.

В настоящее время происходит активное внедрение татарского языка в мобильные устройства. Создаются локализованные сервисные приложения: татарская клавиатура, словари, системы предиктивного набора текста, игры, обучающие программы для различных систем (iOS, Android, Microsoft Windows). С 2016 г. началась татарская локализация операционной системы Аврора для мобильных устройств и планшетов. Она стала первой мобильной операционной системой, дающей возможность полноценно использовать татарский язык наравне с русским языком в мобильных устройствах. Татарская локализация ОС Аврора представляет собой полную адаптацию мобильной операционной системы на этот язык. Общий объем переведенного контента составил порядка 17000 слов.

Еще одна уникальная пользовательская система, адаптированная для татарского языка нашими специалистами совместно с фирмой Смарткат – российская система для профессионального перевода Smartcat (smartcat.ai). Она предназначена для широкого применения в качестве инструмента профессионального переводчика с различными полезными функциями (машинный перевод, электронные словари и т. п.). Эта платформа для автоматизации перевода, которая оптимизирует процесс работы и комплексно решает все переводческие задачи, позволяя создавать проекты, следить за процессом работы команды переводчиков в реальном времени, проверять переведенные сегменты, обсуждать детали с командой прямо в системе. Система Smartcat в настоящее время активно используется в Республике Татарстан государственными органами, учреждениями и муниципальными образованиями. В аккаунте «Официальный Татарстан» зарегистрировано более 160 профессиональных переводчиков из 67 организаций республики.

Терминотворчество является одной из ключевых позиций в процессе локализации компьютерных систем, поскольку именно от того, насколько этот процесс органичен, насколько термины и понятия корректны и адекватны языковым явлениям, зависят цельность и устойчивость языка, а также, что не менее существенно, корректность понимания команд и функций и, соответственно, корректность решения задач пользователя. В настоящее время нередки случаи употребления на практике нескольких татарских вариантов перевода одного и того же английского термина. Например: *санак, компьютер, кампитр, ЭХМ*; *клавиатура – төймәсар, клавиатура; меню – меню, сайлак; монитор – дисплей, күрәк.*

В основу принципов перевода терминов при татарской локализации легли правила образования и применения терминов и понятий в татарском языке, разработанные татарскими филологами [2], а также результаты наших исследований, полученные при локализации

компьютерных систем, при создании англо-татарско-русского толкового словаря по терминам информатики и при обучении студентов [3].

Татарская локализация операционной системы MS Windows решала следующие задачи, требующие перевода терминов и понятий:

1) перевод интерфейсов операционных систем и ее офисных приложений – перевод текстов, которые отображаются на экране дисплея (например, текстов, используемых при работе с электронной почтой, с приложениями Windows Word, Excel и т.д.);

2) перевод на татарский язык текстов на кнопках меню (Save, Open, Close, OK, Yes, Delete, Print и т.д.); перевод файлов справок и помощи на татарский язык.

При этом к татарским текстам, понятиям и терминам предъявлялись следующие требования:

1) правильность, естественность, точность татарских понятий и терминов и татарских текстов для адекватного перевода смысла текста на английском языке;

2) краткость, понятность и правильность понятий, терминов, текстов (команд, действий) на кнопках меню;

3) понятность и компактность понятий, терминов, текстов в файлах справок.

Татарская локализация операционной системы и офисных приложений – это не прямой перевод с английского или русского языков, а творческая адаптация программного продукта для комфортной работы татаро-язычного пользователя в соответствующей операционной среде. Важно, чтобы представленный в системах на разных языках текст имел одинаковый смысл. Тексты на разных языках должны восприниматься одинаково и быть корректными по отношению к пользователю.

К основным принципам образования татарских понятий и терминов, описанным в работе [2], нами предложены дополнительно три новых принципа. Они практически еще не изучены или недостаточно отражены в татарской лексикологии, подробное их описание дается в работе Д. Ш. Сулейманова, А. Ф. Галимянова [1]. Здесь ограничимся определением этих принципов и примерами.

1) Принцип «формального гнезда». Образование новых слов из каркаса (матрицы) татарского слова, в котором «убираются» все гласные буквы, и слово-схема заполняется другими татарскими гласными буквами.

Структуру татарских корневых слов можно представить как матрицу из согласных букв, заполненных гласными буквами. В современном татарском языке 9 гласных букв (*а-ә, о-ө, у-ү, ы-е, и*). Соответственно, заполняя, например, матрицу «*т-з*», можно образовать следующие слова: *таз-тәз-тоз–төз-туз-түз-тыз-тез-тиз*. Здесь 7 слов из 9 представлены в татарском словаре: *таз (тазик), тоз (соль), төз (стройный), туз (береста), түз (терпи), тез (строй), тиз (быстро)*. В то же время по схеме «*җ-л*» образуется всего одно слово, это слово *җил (ветер)*, и нет ни одного татарского литературного слова, которое можно породить по схеме «*б-н*». Такие схемы мы называем «формальными гнездами», это практически готовые структуры для порождения новых слов.

В настоящее время в компьютерных технологиях используется ряд понятий, названия которых представляют собой новые слова, неологизмы, образованные на основе описанного принципа: *күрәк* (*к-р-к*) – дисплей, *күрсәр* (*к-рес-р*) – курсор, *сайлак* (*с-л-к*) – меню, *буяк* (*б-й-к*) – тонер, *турак* (*т-р-к*) – измельчитель бумаги. Возвращенное слово *санак* (компьютер) также могло бы образоваться на основе формы «*с-н-к*».

2) Принцип «возвращенных» слов. Возвращение в язык слова, которое этимологически является тюркским, использовалось в языке, сохранилось в других языках, или вышло из употребления в языке как «архаизм».

Одним из таких «возвращенных» слов является слово айкен (иконка, *icone*), обозначающее пиктограмму – небольшое растровое символическое изображение, используемое в графическом интерфейсе пользователя для выбора того или иного инструмента (программы) или файла.

Слово *айкен-айкөн* составлено из двух тюркских слов *ай – луна* и *кен (көн)* – солнце. У индейцев майя имеется такой глиф с изображением луны и солнца. Соответственно, вполне логично пиктограмму (знак) *иконка* по-татарски называть возвращенным словом *айкен*.

3) Принцип «блендинг» («слияние») – синтетический принцип. Образование новой метафоры из нескольких метафор «слиянием, соединением» и обозначение ее новым именем. Этот принцип в настоящее время активно применяется для языков индоевропейской группы. Например, слово *edutainment (education + entertainment)* – (обучение + развлечение) обозначает программное обеспечение, которое развлекает, когда обучает. Также слиянием слов *пүчтәк (рус.: пустяк, пустячок)* и *күчтәнәч (презент)* можно образовать слово *пүчтәнәч*. В итоге данное новое слово передает новое значение: маленький, недорогой подарок, слово образовано из значений двух метафор. Данный принцип в татарском языке в настоящее время практически не используется для образования новых терминов и понятий.

Программные системы, технологии и лингвистические ресурсы для татарского языка

Следующее направление исследований и разработок института ориентировано на создание программного инструментария, прикладных программ и лингвистических ресурсов для татарского языка, обеспечивающих использование компьютерных систем и технологий в работе с татарским языком во всех сферах и формах его проявления (наука, образование, делопроизводство, издательская деятельность, накопление и обработка информации, и т. д.). Исследования теоретических и прикладных проблем компьютерной лингвистики применительно к татарскому языку являются весьма наукоемкой задачей и требуют интеграции знаний и умений специалистов в смежных областях – информатике, математике и лингвистике. Описанию современного состояния исследований и разработок в этом направлении и раскрытию приведенных ниже разработок в функциональном, содержательном и технологическом аспектах посвящена коллективная монография [4].

В настоящее время ряд задач: создание программных средств и пакетов прикладных программ поддержки татарского языка в инфокоммуникационных технологиях, создание национального корпуса татарского языка и иных лингвистических ресурсов лингвистических ресурсов – решаются в рамках выполнения Государственной программы по изучению, сохранению и развитию языков народов РТ.

Морфологический анализатор татарского языка и программный комплекс снятия морфологической неоднозначности (http://tatmorphan.pythonanywhere.com/).

Как известно, татарский язык обладает содержательно богатой и формально элегантной, регулярной, почти автоматной, морфологией [5]. Морфологическая модель татарского языка является базовой составляющей практически во всех полнофункциональных лингвистических процессорах. Соответственно, создание компьютерной модели морфологии татарского языка было одной из первых и важных задач института. Учитывая структурную специфику татарского языка и исходя из прикладных задач, к настоящему времени нами 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 агглюти-

нативный характер языка, распознавая словоформы потенциально неограниченной длины. *Парадигматическая модель* татарской морфологии обеспечивает быстрое распознавание словоформ и анализ корректности татарских словоформ с точностью до 95% и используется в операционной среде MS Windows и ее офисных приложениях.

Кроме того, в рамках совместного проекта с Белкентским университетом (Турция) разработана *двухуровневая модель морфологии* татарского языка, реализованная в среде программной оболочки PC KIMMO. Эта модель морфологического анализа включена в состав поисковой системы УИС «Россия» (ЦИТ МГУ), причем скорость распознавания составляет 100 слов за 0.014 секунд, что на порядок превосходит требования заказчика. На ее основе разработан морфологический анализатор татарского языка, используемый как средство аннотации текстов, и программный комплекс снятия морфологической многозначности в корпусе татарского языка.

Создана также *гибридная модель морфологического анализа*, использующая генеративный и парадигматический подходы и являющаяся частью информационно-инструментального комплекса «Татарская морфема» [6].

В татарском языке, как и в других агглютинативных языках тюркской группы, морфемы представляют собой важнейшие значащие языковые единицы, которые несут как семантическую, так и синтаксическую информацию. Имея теоретически неограниченное количество присоединяемых к основе морфем, *морфологическая многозначность* приобретает разнообразные формы, что значительно усложняет задачу разрешения. В связи с этим весьма важным и значимым теоретическим и практическим результатом является решение следующих задач, успешно выполненных в институте: 1) разработка программного инструментария для автоматизации процессов создания и тестирования программных средств разрешения многозначности в ЕЯ-текстах; 2) создание лексико-грамматической модели представления базы знаний для разрешения морфологической многозначности в ЕЯ-текстах; 3) разработка программных средств разрешения морфологической многозначности в ЕЯ-текстах на татарском языке.

Система машинного перевода (СМП) «Татсофт» (translate.tatar). Первый нейросетевой русско-татарский машинный переводчик. Один из важных и востребованных продуктов. В настоящее время является одним из лучших по качеству перевода среди своих аналогов (Яндекс и Google). По данным Яндекс.Метрики по состоянию на 30.08.2023 г. количество стран, обратившихся за сервисом, составило 102 страны, обработано более 20 миллионов запросов на перевод. В «Татсофте» перевод осуществляется с «пониманием» всего предложения, а не отдельных слов/фраз. Модели переводчика обучены на базе параллельных текстов, которая включает более 1 млн русско-татарских пар предложений и двуязычные специализированные словари фамилий, имен, отчеств, государств, субъектов РФ, районов РТ, населенных пунктов, гражданств, национальностей в объеме более 95 тыс. словарных пар.

В настоящее время ведутся работы по разработке речевого (голосового) интерфейса, что сделает сервис машинного перевода «Татсофт» еще более удобным и особенно важным для пользователей со слабым зрением.

В институте разработана также система машинного перевода для 7 языковых пар, в которых один язык является русским, а другой принадлежит к тюркской языковой группе. Для достижения поставленной цели создана база параллельных обучающих данных и разработан метод объединения собранных параллельных корпусов на основе структурно-функциональной модели тюркских морфем [6], а также созданы программные средства для обучения многоязычного машинного переводчика на основе подходов трансферного

обучения и увеличения данных. Это позволило впервые создать параллельный корпус для крымскотатарско-русской языковой пары. Система машинного перевода работает еще с 6 языковыми парами (татарско-русский, башкирско-русский, чувашско-русский, казахско-русский, кыргызско-русский и узбекско-русский).

Татарский национальный корпус (ТНК) «Туган тел» (http://tugantel.tatar/). Очевидно, формирование максимально полной, репрезентативной лингвистической ресурсной базы также служит сохранению, изучению и развитию языка. В этом плане одним из важных разработок института является Татарский корпус «Туган тел» («Родной язык»). Это лингвистический ресурс современного литературного татарского языка, адресованный широкому кругу пользователей: лингвистам, специалистам в области татарского, тюркского и общего языкознания, типологам, преподавателям татарского языка, деятелям культуры, а также всем, кто изучает татарский язык и интересуется им. В настоящее время объем корпуса составляет около 200 миллионов словоупотреблений и содержит тексты различных жанров (художественная литература, тексты СМИ, тексты официальных документов, учебная литература, научные публикации и др.). Уникальность корпуса заключается еще и в том, что программная платформа корпуса языка представляет собой оригинальную разработку нашего института и имеет инструментарий, практически «заточенный» под татарский язык.

Проект выполнялся в рамках Государственной программы «Сохранение, изучение и развитие государственных языков Республики Татарстан и других языков в Республике Татарстан на 2014–2020 годы». Практически данный ресурс реализует базовые блоки концепции машинного фонда татарского языка МФТЯ, представленной в статье [7] средствами современных технологий.

Электронная версия Атласа татарских народных говоров **(**atlas.antat.ru)**.** Еще один уникальный продукт: Электронный Атлас татарских народных говоров. Атлас построен совместными усилиями специалистов НИИ «Прикладная семиотика» АН РТ, ИЯЛИ АН РТ и КФУ, охватывает все основные районы расселения татар и отражает сведения по фонетике, морфологии, лексике и синтаксису татарского языка, собранные в 28 регионах Российской Федерации. Выпуск электронного варианта атласа является новым этапом представления диалектических знаний по татарскому языку на базе геоинформационных систем. Реализована специальная программа для сбора материалов и включения их в базу диалектологического Атласа силами пользователей. Электронный атлас татарских говоров позволяет получить информацию об имеющихся языковых явлениях в привязке к конкретным географическим объектам на картах. На данный момент используется информация с 215 карт языковых явлений.

Интернет-платформа «Тюркская морфема». Еще одна разработка – интернет-платформа «Тюркская морфема» [6] – это инструмент, разработанный с целью создания интегрального описания тюркских языков в рамках целостной лингвистической модели. Подход создания Интегральной лингвистической модели естественного языка был предложен Мельчуком при работе над проектом машинного перевода «Этап» [8]. В модели, использованной в проекте «Этап», в качестве базовой лингвистической единицы была взята лексема, которой в явном виде прописывали все ее свойства, включая просодические, семантические, прагматические, коммуникативные и сочетаемостные. В модели портала «Тюркская морфема» в качестве базовой единицы использована тюркская морфема, как аффиксальная, так и корневая. База данных модели представляет собой единый граф знаний, содержащий с описанием характеристик тюркских морфем на всех языковых уровнях (фонологическом, морфологическом, синтаксическом, семантическом). Основные функ-

ции сервиса: формирование ресурсной базы для программных продуктов, осуществляющих компьютерную обработку тюркских языков, таких как системы машинного перевода, информационно-поисковые системы, системы разметки электронных корпусов, извлечения данных и др.; информационно-справочная система, содержащая практически полную информацию о тюркских морфемах; инструментарий для исследований ученых-тюркологов.

Важно отметить, что эта интернет-платформа есть одновременно и ценный программный инструментарий, а также лингвистическая база для сравнительного изучения тюркских языков и реализации совместных коллективных проектов с участием ученых, являющихся носителями языка этноса, с глубокой лингвистической интуицией.

Система анализа и синтеза татарской речи. Как прогнозируется специалистами, одним из основных направлений развития в сфере высоких технологий в ближайшие годы будут речевые технологии, особенно, автоматическое распознавание речи (ASR). Ожидается широкое внедрение технологий ASR в ведущие сектора экономики. В Институте разрабатывается комплекс речевых технологий, включающих в себя возможности определения языка говорящего, синтеза и распознавания татарской речи [9]. Накапливаются и анализируются базы данных текстовой и речевой информации на татарском языке, разрабатываются технологии машинного обучения, происходит интеграция речевого интерфейса на татарском языке в современные ПК и мобильные устройства. Достигнутые результаты сравнимы с мировыми аналогами и позволяют организовать «общение» с компьютером с помощью голосовых команд (речевой перевод, мобильные ассистенты, диктовка сообщений, чтение новостей).

Заключение

В статье изложены результаты деятельности НИИ «Прикладная семиотика» Академии наук РТ за последние годы в области создания программных систем, технологий и лингвистических ресурсов с целью обеспечения паритетного функционирования татарского языка в инфокоммуникационных технологиях в качестве одного из государственных языков в Республике Татарстан. Разработанные институтом и представленные в статье программные продукты способствуют сохранению и развитию татарского языка, повышению его активности в интернете и помогают татарскому языку стать языком компьютерных технологий. Описан ряд потенциальных когнитивных возможностей татарского языка, позволяющих ему стать формальной базой для построения новых средств описания, хранения и обработки информации.

Литература

1. Сулейманов, Д. Ш. Система татарских терминов в компьютерных технологиях и информатике / Д. Ш. Сулейманов, А. Ф. Галимянов // В сб. трудов Первой международной конференции «Компьютерная обработка тюркских языков». – Астана : ЕНУ им. Л. Н. Гумилева, 2013. – С. 132–140.

2. Правила образования, совершенствования и применения татарских терминов. Комитет при Кабинете Министров РТ по реализации Закона РТ «О языках народов РТ». Зам. председателя Комитета профессор М. З. Закиев, председатель терминологической комиссии Комитета, доцент И. М. Низамов. – Казань, 1995. – 13 с. (на татарском языке).

3. Сулейманов, Д. Ш. Англо-русско-татарско-чувашский словарь терминов по информатике и информационным технологиям (с толкованиями на татарском языке). [Текст] / Д. Ш. Сулейманов, А. Ф. Галимянов, М. Х. Валиев [и др.]. // Приложение к Материалам III Международной конференции по компьютерной обработке тюркских языков (TurkILang 2015, Казань, 17–19 сентября 2015 г.). – Казань : Издательство АН РТ, 2015. – 400 с.

4. Формальные модели и программные инструменты компьютерной обработки татарского языка / Р. Р. Гатауллин, А. Р. Гатиатуллин, Р. А. Гильмуллин [и др.] ; под редакцией Д. Ш. Сулейманова,

А. Ф. Хусаинова ; Академия наук РТ, Институт прикладной семиотики АН РТ. – Казань : Издательство Академии наук РТ, 2019. – 260 с.

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

6. Ayrat Gatiatullin, Dzhavdet Suleymanov, Nikolai Prokopyev, Bulat Khakimov (2020). About Turkic Morpheme Portal. *In Proceedings of the Computational Models in Language and Speech Workshop* (CMLS 2020). Kazan, Russian, November 12–13, 2020. Pp. 226–243 (http://ceur-ws.org/Vol-2780/).

7. Бухараев, Р. Г. Машинный фонд татарского языка : состояние и проблемы / Р. Г. Бухараев, Д. Ш. Сулейманов // Инфокоммуникационные технологии глобального информационного общества : тезисы докладов научно-практической конференции (Казань, 16–18 сентября 2003). – Казань : Издательство КГУ, 2003. – С. 80–81.

8. Мельчук, И. А. Опыт теории лингвистических моделей «Смысл –Текст» / И. А. Мельчук. – Москва, 1974.

9. Khusainov A., Suleymanov D. (2015). An approach to automate process of creating speech analysis systems for under-resourced languages. IEEE CPS Volume of Proc. of MICAI-2015. (Cuernavaca, October 25 to 31, 2015). 2015. P. 28-34. [Tatar National Corpus] Tatar National Corpus – URL: http://tugantel.tatar/.

References

1. Sulejmanov, D. SH. Sistema tatarskih terminov v komp'yuternyh tekhnologiyah i informatike / D. SH. Sulejmanov, A. F. Galimyanov // V sb. trudov Pervoj mezhdunarodnoj konferencii «Komp'yuternaya obrabotka tyurkskih yazykov». – Astana : ENU im. L. N. Gumileva, 2013. – S. 132–140.

2. Pravila obrazovaniya, sovershenstvovaniya i primeneniya tatarskih terminov. Komitet pri Kabinete Ministrov RT po realizacii Zakona RT «O yazykah narodov RT». Zam. predsedatelya Komiteta professor M. Z. Zakiev, predsedatel' terminologicheskoj komissii Komiteta, docent I. M. Nizamov. – Kazan', 1995. – 13 s. (na tatarskom yazyke).

3. Sulejmanov, D. SH. Anglo-russko-tatarsko-chuvashskij slovar' terminov po informatike i informacionnym tekhnologiyam (s tolkovaniyami na tatarskom yazyke). [Tekst] / D. SH. Sulejmanov, A. F. Galimyanov, M. H. Valiev [i dr.]. // Prilozhenie k Materialam III Mezhdunarodnoj konferencii po komp'yuternoj obrabotke tyurkskih yazykov (TurkILang 2015, Kazan', 17–19 sentyabrya 2015 g.). – Kazan' : Izdatel'stvo AN RT, 2015. – 400 s.

4. Formal'nye modeli i programmnye instrumenty komp'yuternoj obrabotki tatarskogo yazyka / R. R. Gataullin, A. R. Gatiatullin, R. A. Gil'mullin [i dr.] ; pod redakciej D. SH. Sulejmanova, A. F. Husainova ; Akademiya nauk RT, Institut prikladnoj semiotiki AN RT. – Kazan' : Izdatel'stvo Akademii nauk RT, 2019. – 260 s.

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

6. Ayrat Gatiatullin, Dzhavdet Suleymanov, Nikolai Prokopyev, Bulat Khakimov (2020). About Turkic Morpheme Portal. In Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020). Kazan, Russian, November 12–13, 2020. Pp. 226–243 (http://ceur-ws.org/Vol-2780/).

7. Buharaev, R. G. Mashinnyj fond tatarskogo yazyka : sostoyanie i problemy / R. G. Buharaev, D. SH. Sulejmanov // Infokommunikacionnye tekhnologii global'nogo informacionnogo obshchestva : tezisy dokladov nauchno-prakticheskoj konferencii (Kazan', 16–18 sentyabrya 2003). – Kazan' : Izdatel'stvo KGU, 2003. – S. 80–81.

8. Mel'chuk, I. A. Opyt teorii lingvisticheskih modelej «Smysl –Tekst» / I. A. Mel'chuk. – Moskva, 1974.

9. Khusainov A., Suleymanov D. (2015). An approach to automate process of creating speech analysis systems for under-resourced languages. IEEE CPS Volume of Proc. of MICAI-2015. (Cuernavaca, October 25 to 31, 2015). 2015. P. 28-34. [Tatar National Corpus] Tatar National Corpus – URL: http://tugantel.tatar/