ОСНОВНЫЕ КОМПОНЕНТЫ ПРИКЛАДНОЙ ГРАММАТИЧЕСКОЙ МОДЕЛИ ТАТАРСКОГО ЯЗЫКА[1]

MAIN COMPONENTS OF APPLIED GRAMMATICAL MODEL OF TATAR LANGUAGE

Сулейманов Д.Ш. *(dvdt@telecet.ru), Казанский государственный университет*
Невзорова О.А. *(olga.nevzorova@ksu.ru),*
Татарский государственный гуманитарно-педагогический университет, НИИММ им. Н.Г. Чеботарева
Гатиатуллин А.Р. *(Agat1972@mail.ru), Академия наук РТ*
Гильмуллин Р.А. *(abbyy_turkish@mail.ru), Казанский государственный университет*
Аюпов М.М., *Казанский государственный университет*
Пяткин Н.В. *(nikolaip@mail.ru), НИИММ им. Н.Г. Чеботарева, Казань*

В статье представлены описания основных компонент прикладной грамматической модели татарского языка для задач информационного поиска.

1. Введение

Технологическая задача внедрения татарского языка в современные компьютерные технологии в настоящее время успешно решена. В рамках реализации практических задач фундаментальной программы Академии наук Республики Татарстан “Компьютерное обеспечение функционирования татарского языка как государственного. Концептуально-алгоритмическая модель татарского языка” разработаны экранные и клавиатурные драйверы, драйверы печати и шрифтовое обеспечение для татарского языка на кириллической основе. На основе принятых стандартов кодировки символов татарского алфавита по соглашению с фирмой Microsoft были разработаны и внедрены в операционную среду Windows NT 5.0 и Office-2000 модули поддержки татарской локализации. Очередным важным шагом является полная татарская локализация операционной системы Windows XP и Windows Vista.

Однако для эффективного встраивания татарского языка в различные информационные технологии, в том числе технологии информационного поиска, необходима разработка прикладной грамматической модели татарского языка. Прикладное назначение определяет состав модели, включающий морфологическую и синтаксическую компоненту. На рис. 1 представлены основные компоненты прикладной грамматической модели татарского языка и реализуемые на их основе функциональные возможности для поиска. Лексический поисковый индекс строится на основе двухуровневой автоматной модели морфологии; построение сложного индекса, включающего различные аналитические конструкции, осуществляется с привлечением частичного синтаксического анализа.

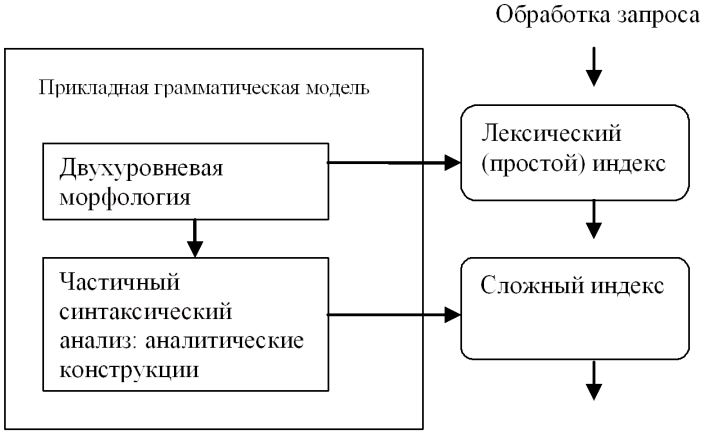

Рис.1. Состав и назначение прикладной грамматической модели

[1] Работа выполнена при частичной финансовой поддержке РФФИ (грант № 04-06-97501) и Инвестиционно-венчурного фонда РТ.

Авторский коллектив имеет более чем десятилетний опыт разработки языковых моделей и технологий использования лингвистических моделей в современных компьютерных технологиях. К числу фундаментальных результатов относятся разработки морфологического корректора татарских текстов, двухуровневой модели татарской морфологии, экспериментальной версии синтезатора татарской речи, интегрированного программно-информационного комплекса “Татарская морфема”, машинного фонда татарского языка.

Разрабатываемые лингвистические модели относятся к классу концептуально-функциональных моделей [1], по сути являющимися структурно-функциональными моделями определенных языковых уровней. Так, программно-информационный комплекс “Татарская морфема” [2] представляет собой реализацию структурно-функциональной модели татарских морфем. Татарские морфемы исследуются и описываются на всех языковых уровнях, т.е. морфологии, синтаксиса и семантики. Разработка прикладной грамматической модели татарского языка представляет собой дальнейшее развитие направления структурно-функционального моделирования и связана с построением полной морфологической модели татарского языка, а также с разработкой ряда специальных синтаксических моделей, прежде всего моделей аналитических конструкций.

2. Задача информационного поиска в татарско-русской электронной коллекции текстов

В настоящее время весьма актуальным является разработка специализированных текстовых баз данных, в том числе содержащих документы на разных языках, и обеспечение многоязыкового поиска. Однако в настоящее время отсутствуют системы подобного класса, предназначенные для обработки коллекций текстов на татарском и русском языках. Главная задача настоящей статьи – описать основные компоненты прикладной грамматической модели татарского языка, разработанной для внедрения в поисковые технологии.

Современные поисковые технологии позволяют адаптировать основные поисковые механизмы для любого нового языка, для которого разработана соответствующая лингвистическая поддержка на уровне развитых морфологических и частичных синтаксических моделей. Такой уровень поддержки татарского языка практически полностью обеспечивается функциональностью прикладной грамматической модели татарского языка.

Одной из известных информационных поисковых систем, осуществляющих полнотекстовый поиск, является Университетская информационная система РОССИЯ (НИВЦ МГУ) [4]. Интегрирование прикладной грамматической модели татарского языка в УИС РОССИЯ позволяет эффективно поддерживать многоязычный поиск в татарско-русской электронной коллекции текстов.

Функциональные возможности УИС РОССИЯ позволяют:

- вести большие полнотекстовые базы данных для интеграции общественно-политической информации о жизни региона и федерации в целом;
- реализовывать стандартные возможности полнотекстового поиска – контекстный поиск по документам базы данных с учетом морфологии, включая подсветку результатов поиска;
- автоматически выделять формальную метаинформацию для обрабатываемых документов – автор, заглавие, дата и т.п. Язык запросов позволяет включать в запрос условия по любой метаинформации одновременно для нескольких классов документов.

Аналитические возможности УИС РОССИЯ позволяют автоматически выделять тематическую метаинформацию (построение терминологического индекса по общественно-политическому тезаурусу, автоматическая рубрикация одновременно по нескольким рубрикаторам, автоматическое аннотирование). Уникальной особенностью является возможность качественной автоматической рубрикации по иерархическим рубрикаторам большого размера (более 500 рубрик). Язык запросов позволяет включать в запрос логические условия любой сложности, где частным условием является любой элемент метаинформации или контекста.

УИС РОССИЯ поддерживает обработку многоязыковых текстов [5]. При этом под многоязычием (подключением другого языка) понимается:

- возможность построения запросов на различных языках к текстам коллекции;
- возможность использования многоязычный тезаурус по общественно-политической тематике для обработки и поиска документов;
- возможность подсветки результатов запроса для обоснования релевантности документа;
- возможность анализа содержания иноязычного документа средствами на родном языке.

Возможность построения запроса на татарском языке к русско-татарской коллекции текстов базируется на морфологии разбора документа и построении морфологического поискового индекса. Многоязычный тезаурус по общественно-политической тематике для обработки и поиска документов, снабженный синонимическими рядами концептов на татарском языке, обеспечивает механизмы смены языка запросов.

3. Компоненты прикладной грамматической модели татарского языка: двухуровневая морфологическая модель

Важнейшей компонентой прикладной грамматической модели татарского языка является морфологическая модель. В качестве формальной модели для построения модели татарской морфологии была выбрана двухуровневая модель автоматов с конечными состояниями, реализованная в среде программного инструментария PC-KIMMO [3].

Технология автоматов с конечными состояниями для автоматического распознавания и генерации словоформ применяется с начала 80-х годов. Она основана на представлении, что правила морфологических альтернатив могут быть реализованы автоматами с конечными состояниями. Также известно, что возможные комбинации основ и аффиксов могут быть кодированы как сеть автоматов с конечными состояниями. Двухуровневый распознаватель, используя преобразователи, отображает поверхностную строку в последовательность ветвей в дереве букв и вычисляет основу, исходя из информации, имеющейся в границах ветвей.

Информационная база двухуровневого морфологического анализатора (ДМА), с точки зрения разработчика морфологического анализатора, состоит из двух файлов, создаваемых пользователем.

Первый файл - файл правил (Rules) описывает алфавит и фонологические правила. Второй файл – лексикон, содержит словарь лексических единиц (корневых и аффиксальных морфем) и их толкования, а также описания морфотактических правил.

При двухуровневом подходе фонология определяется как связь между лексическим уровнем глубинного представления слов и их реализации на поверхностном уровне, в силу чего теоретическая модель фонологии ДМА называется двухуровневой фонологией. ДМА включает две функциональные компоненты - генератор и распознаватель.

Генератор на входе получает лексическую форму, применяет правила фонологии и возвращает соответствующую поверхностную форму. На этом этапе лексикон не используется. Распознаватель получает на входе поверхностную форму, применяет правила фонологии, обращается к лексикону и возвращает соответствующие лексические формы с их грамматическими характеристиками (толкованиями). В приводимых ниже примерах элементы глубинного представления кодируются строками из заглавных букв (*-ДА*), а элементы поверхностного уровня – строками из малых букв (*-да*). Элементам глубинного представления могут соответствовать различные поверхностные реализации (алломорфы).

На рис. 2 представлена структурно-функциональная схема ДМА татарского языка с некоторыми примерами входных и выходных данных. Так, генератор, используя файл фонологических правил, лексический вход *урман+ДА* переводит в поверхностную форму – *урманда.* Распознаватель, используя файлы фонологических и морфотактических правил, раскладывает входную словоформу (поверхностную форму) *урманда* по составляющим, выдавая соответствующие содержательные описания: **Noun** (*урман*)+[**Case_Loc** (*ДА*)], где **Noun** – грамматический класс существительных, **Case_Loc** – класс падежных аффиксов.

Оболочка ДМА обеспечивает разработку, тестирование и отладку двухуровневых описаний.

Автоматная модель грамматики татарского языка описывается конечным списком множеств, задающих

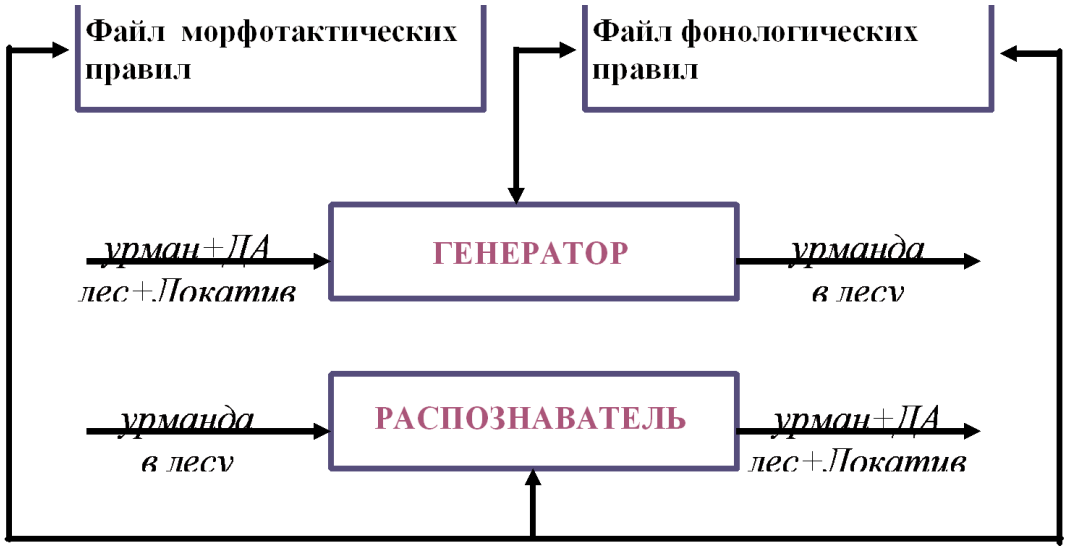

Рис. 2. Структурно-функциональная схема ДМА татарского языка

алфавит и подмножества фонологических описаний языка, а также множествами фонологических и морфотактических правил. Важной составляющей фонологического описания является множество парных символьных соответствий лексического и поверхностного уровней. Ядро фонологического описания составляет множество фонологических правил, устанавливающих соответствие символов в зависимости от контекста, т.е. символьного окружения в словоформе. Файл фонологических правил татарского языка содержит 39 правил. Файл морфотактических правил разработан на основе морфотактических схем и определяет взаимосвязи между основой и аффиксальными группами. Лексикон корневых лексем построен на основе современного татарского языка и состоит из ряда лексиконов, заполненных согласно соответствующим требованиям ДМА.

Подлексиконы содержат строки лексических входов, состоящие из трех частей:

1) Лексический атом (татарское корневое слово).

2) Класс (подлексикон) присоединения; т.е. единицы, которые могут присоединяться непосредственно справа. Классы присоединений могут следовать за различными морфемными единицами. Например, лексикон ALTERNATION в ДМА определяет список имен подлексиконов в порядке их следования в словоформе.

3) Описание грамматических признаков. Как правило, в этом разделе описываются морфологические, грамматические, лексические, или семантические свойства лексической единицы.

При обработке слова распознавателем описания каждой морфемы добавляются в строку результата.

Ниже приводится описание фрагмента файла морфотактических правил для татарского глагола с примерами и комментариями.

ALTERNATION BEGIN VERBS NOUN ADJECTIVE NUMERAL PRONOUN ADVERB SPECIAL {означает, что в данном описании определено, что имеется 7 разных возможностей для начала татарского слова);

VERB - подлексикон для глаголов;

NOUN - подлексикон для существительных;

ADJECTIVE - подлексикон для прилагательных;

NUMERAL - подлексикон для числительных;

PRONOUN - подлексикон для местоимений, послелогов;

ADVERB - подлексикон для наречий;

SPECIAL - подлексикон для союзов, междометий.

ALTERNATION **verb** REFLEX MODAL CAUS NOMINATIVE INFINITIVE PARTICIPAL CONTRARY IMPERATIVE REQUEST {указаны аффиксальные классы, которые могут следовать за глаголом. В нашем случае - это указанные 11 аффиксальных классов, каждый из которых доопределяется далее вплоть до соответствующей группы аффиксов}

LEXICON VERB // список глагольных основ

бар verb “V(bar)“

кил verb “V(kil)“

кара verb “V(kara)“

.....

Далее следует описание фрагмента аффиксальной базы глагольных словоформ татарского языка.

В первой части описания лексикона приводится аффиксальная морфема, затем название класса морфем, который может следовать за этим аффиксом. Третья составляющая содержит грамматические описания данного лексического входа.

Пример описания подлексикона для представления отрицательной формы глагола:

LEXICON CONTRARY // список отрицательных форм для глагола

+мА IMPERATIVE «+NEG(мА)»

+мА CONDITIONAL «+NEG(мА)»

+мА TENSE «+NEG(мА)»

+мА REQUEST «+NEG(мА)»

Пусть, на вход анализатору подаются следующие слова: *барма (не ходи), barmaca (если не пойдет), бармады (он не ходил), бармадымы (не ходил?).*

Результат работы модуля распознавания:

бармы [V(бар)+NEG(мА)]

бармаса [V(бар)+NEG(мА)+COND_3PS_Sing(сА)]

бармады [V(бар)+NEG(мА)+PAST_DEF(ДЫ)]

бармадымы [V(бар)+NEG(мА)+PAST_DEF(ДЫ)+Q(мЫ)]

где V(бар) – глагол в форме ‘*иди’*, NEG(mE) – отрицательный аффикс “мА”, PAST_DEF(ДЫ) – аффикс прошедшего времени “ДЫ”, Q(ДЫ) – аффикс вопросительной формы “мЫ”.

4. Компоненты прикладной грамматической модели татарского языка: модели аналитических конструкций

Построение лексического поискового индекса для татарских текстов имеет ряд существенных особенностей по сравнению с построением аналогичного индекса для русских текстов. В силу специфики татарского языка для эффективного поиска и обеспечения релевантности результата поисковому образу в поисковый индекс необходимо включить не только отдельные лексемы, но и аналитические конструкции различных типов.

Необходимость создания сложного поискового индекса оправдана по следующим причинам:

- особенностью татарского языка является частотное употребление аналитических конструкций для кодирования лексических значений;
- можно определить наиболее частотное конечное ядро лексем и аналитических конструкций, описание которых на начальном этапе исследования позволяет значительно минимизировать объем найденной информации по поисковому запросу.

Таким образом, предлагается подход, учитывающий специфику татарского языка, ориентированный на перенос сложностных (временных и емкостных) аспектов поиска информации на стадию построения индексной базы, в отличие от подходов, использующих сложные фильтры на этапах поиска и анализа результатов поиска.

К множеству аналитических конструкций традиционно относят синтаксические конструкции различных типов: идиомы (*бить баклуши*, *пить горькую*, *водить за нос*); пословицы (*тише едешь – дальше будешь*, *не в свои сани не садись*); поговорки (*вот тебе*, *бабушка*, *и юрьев день*; *лед тронулся*!); фразеосхемы (*X он и в Африке X*; *всем X-ам X*) и др. Для описания аналитических конструкций в татарском языке нами используется структурно-функциональная модель, наиболее полно описывающая свойства языковой единицы на каждом из языковых уровней: морфонологическом, морфологическом, синтаксическом и семантическом.

В настоящее время разрабатываются модели аналитических конструкций в татарском языке следующих четырех типов:

1. Послеложные конструкции: *урманга кадәр ‘до леса’;*
2. Сложные глаголы (существительное + глагол): *колак салу ‘прислушаться’, дәвам итү ‘продолжить’;*
3. Составные глаголы (глагол + глагол): *эшли иде ‘работал’*;
4. Составные имена (существительное + существительное): *үги ана яфрагы ‘мать-и-мачеха’, эт шомырты ‘собачья черемуха’.*

Для каждого типа аналитических конструкций предложена структурно-функциональная модель описания. Так, структурно-функциональная модель сложных глаголов, состоящих из двух словоформ, имеет следующие разделы:

1. Лексический аспект
    - 1.1. Основа первой словоформы
    - 1.2. Часть речи первой словоформы
    - 1.3. Основа второй словоформы
    - 1.4. Часть речи второй словоформы
2. Морфологический аспект
    - 2.1. 1-й обязательный аффикс первого слова
    - 2.2. 2-й обязательный аффикс первого слова.
    - 2.3. Возможные аффиксы у первого слова
3. Синтаксический аспект
    - 3.1. Конструкция в виде предложения
    - 3.2. Конструкция в виде словосочетания
    - 3.3. Возможность слов между первым и вторым словом
4. Семантический аспект
    - 3.1. Буквальное значение
    - 3.2. Идиоматическое значение
    - 3.3. Синонимия
    - 3.4. Антонимия

Пример фрагмента заполнения параметров структуры для сложного глагола *хәтергә төшерү ‘напомнить, вспомнить’*:

1. Лексический аспект
    - 1.1. Основа 1-го слова: *хәтер ‘память’*
    - 1.2. Часть речи: имя существительное
    - 1.3. Основа 2го слова: *төшер ‘опусти’*

1.4. Часть речи: глагол
2. Морфологический аспект
2.1. 1-й аффикс: *-ГА*
2.2. 2-й аффикс: нет
2.3. Возможные аффиксы у первого слова: *-Лар, -Ым, Ың, -[с]Ы, ЫбЫз, ЫгЫз.*

Для современного татарского языкознания подобные структурно-функциональные описания аналитических конструкций важны не только в теоретическом плане, как интегральные описания языковых структур, но и имеют реальные перспективы использования в прикладных разработках. Одним из примеров практического приложения моделей аналитических конструкций является их использование в поисковой системе для построения сложного поискового индекса.

Для поддержки поиска достаточно ограниченных описаний структурно-функциональной модели, включающих лексический и морфологический аспекты. Такая база данных аналитических конструкций разработана для внедрения в УИС РОССИЯ. В настоящее время заполнена база данных двухсоставных аналитических конструкций 4-х типов, каждый из которых содержится в отдельной таблице (словаре).

Разработаны словари следующих типов: послеложные конструкции, сложные глаголы, составные глаголы и составные имена.

5. Заключение

В статье описаны основные компоненты прикладной грамматической модели татарского языка, необходимые для функционирования приложений информационного поиска в многоязычных корпусах текстов. Основными компонентами прикладной модели являются двухуровневый морфологический анализатор татарского языка и прагматически-ориентированная модель аналитических конструкций. Все полученные результаты являются новыми и оригинальными.

Тестирование морфологической модели татарского языка проводилось на подготовленной авторами двуязычной (татарско-русской) электронной коллекции текстов по общественно-политической тематике. Подготовка коллекции вскрыла ряд серьезных организационно-правовых проблем, связанных с охраной авторских прав, отсутствием норм для организации электронных хранилищ некоммерческого использования, ряд технологических проблем представления текстов электронной коллекции.

Список литературы

1. Сулейманов Д.Ш. Обработка ЕЯ-текстов на основе прагматически-ориентированных лингвистических моделей // Обработка текста и когнитивные технологии. Вып.3., 1998. С.205-212.
2. Сулейманов Д.Ш., Гатиатуллин А.Р. Структурно-функциональная модель татарских морфем. Казань, 2003. 212 с.
3. Antworth E.L. PC-KIMMO: a two-level processor for morphological analysis. Technical Report Occasional Publications in Academic Computing, № 16. Summer Institute of Linguistics, Dallas, Texas, 1994.
4. Агеев М.С., Добров Б.В., Журавлев С.В., Лукашевич Н.В., Сидоров А.В., Юдина Т.Н. Технологические аспекты организации доступа к разнородным информационным ресурсам в университетской информационной системе РОССИЯ // Электронные библиотеки, 2002. Т. 5. Выпуск 2.
5. Добров Б.В., Лукашевич Н.В. Организация двуязычного поиска в Университетской системе РОССИЯ // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Четвертой Всероссийской научной конференции RCDL’2002 (Дубна, 15-17 октября 2002 г.): В 2 т. Дубна: ОИЯИ, 2002. Т.2. С. 148-158.