

ФОРМАЛИЗАЦИЯ И ПОСТРОЕНИЕ СИСТЕМЫ МОРФОЛОГИЧЕСКОГО АНАЛИЗА КРЫМСКОТАТАРСКОГО ЯЗЫКА

© Адживелиева З.Д., Анафиев А.С., Заирова С.И.

ТАВРИЧЕСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ ИМ. В.И. ВЕРНАДСКОГО
ФАКУЛЬТЕТ МАТЕМАТИКИ И ИНФОРМАТИКИ
ПР-Т ВЕРНАДСКОГО, 4, Г. СИМФЕРОПОЛЬ, 95007, УКРАИНА
E-MAIL: anaftiyev@gmail.com

Abstract. The basic features of Crimean-Tatar Language morphology for the purpose of construction the morphological analysis subsystem for the Crimean-Tatar Language understanding system are considered in this article.

ВВЕДЕНИЕ

Автоматическая обработка естественно-языковых (ЕЯ) текстов – одно из приоритетных направлений в области новых информационных технологий. Область применения систем анализа ЕЯ-текстов довольно разнообразна. Сюда можно отнести поисковые системы, вопросно-ответные системы, системы автоматического машинного перевода.

В связи с изменением в последнее время демографической ситуации в Крыму, появлением крымскотатарских школ, ВУЗов, объектов культуры, с учетом развития крымскотатарской науки и искусства, а также для популяризации крымскотатарского языка и культуры представляется целесообразным приступить к созданию моделей и информационных систем понимания крымскотатарского языка и автоматического перевода с крымскотатарского и на крымскотатарский язык. В частности, «крымскотатарско-украинский-крымскотатарский» и «крымскотатарско-русский-крымскотатарский» системы автоматического перевода.

Каждый язык на земле, это наибольшая ценность, которую нужно не только оберегать, но и развивать в правильном направлении для ее сохранения в современном мире.

Целью данной и последующих работ в этом направлении является построение математико-информационной модели понимания крымскотатарского языка с целью дальнейшего применения её в информационных системах регионального управления для автоматического перевода документов из крымскотатарского и на крымскотатарский язык.

1. ЛИНГВИСТИЧЕСКИЙ ПРОЦЕССОР

Как известно ядром любой естественно-языковой системы является лингвистический процессор – компонента системы, реализующая формальную лингвистическую модель и способная работать с ЕЯ во всем его объеме [1].

Основными функциями лингвистического процессора (ЛП) являются:

1. моделирования понимания (анализ);
2. моделирования производства текстов (синтез).

В ЛП различают три уровня пофразного представления текста – морфологический, синтаксический и семантический. Каждому уровню соответствует свой формальный образ – структура. Таким образом, выделяют морфологическую, синтаксическую и семантическую структуру предложения.

Под *морфологической структурой* понимают последовательность слов, входящих в анализируемое предложение, с указанием части речи и морфологических характеристик (падежа, числа, рода, одушевленности, вида и т.п.); под *синтаксической структурой* – дерево зависимостей, в узлах которого стоят слова данного естественного языка с указанием части речи и грамматических характеристик, а дуги соответствуют специфичным для данного естественного языка отношениям синтаксического подчинения; под *семантической структурой* – дерево зависимостей, в вершинах которого стоят либо предметные имена, либо слова универсального семантического языка, а ребра соответствуют универсальным отношениям семантического подчинения, таким, как аргументное, атрибутивное, конъюнкция, дизъюнкция, равенство, неравенство, больше, меньше, принадлежит и т.п. Для семантической структуры существенной информацией является информация о кореферентности узлов – информация о том, в каких случаях речь идет об одном и том же объекте, а в каких – о разных.

В целом лингвистический процессор должен обеспечивать выполнение следующих преобразований:

предложение на ЕЯ → *МорфСтрукт* → *СинтСтрукт* → *СемСтрукт* (анализ);
СемСтрукт → *СинтСтрукт* → *МорфСтрукт* → *предложение на ЕЯ* (синтез),

где *МорфСтрукт* – морфологическая, *СинтСтрукт* – синтаксическая и *СемСтрукт* – семантическая структуры предложения.

Таким образом, для построения ЛП, необходимо разработать:

1. формальные языки для записи (образов) предложений на морфологическом, синтаксическом, семантическом уровнях представления;
2. формальное понятие структуры предложения для каждого из этих уровней;
3. массивы правил для преобразования структур смежных уровней друг в друга;
4. морфологический, синтаксический и семантический словари, включив в них всю информацию о каждой лексеме, необходимую для осуществления соответствующего преобразования.

2. МОРФОЛОГИЧЕСКИЙ АНАЛИЗ КРЫМСКОТАТАРСКОГО ЯЗЫКА

Стадия морфологического анализа (МА) является наиболее проработанным лингвистическим этапом процесса обработки естественного текста. Целью МА является определение морфологической информации словоформ для использования ее на последующих этапах обработки ЕЯ текста.

Крымскотатарский язык является агглютинативным [2], т.е. словоформы образуются только при помощи аффиксов (суффиксов), что делает, учитывая правила следования (порядок в слове) аффиксов различного типа, интересным построение

иерархии классов аффиксов, которую будет удобно использовать для программной реализации морфологического анализатора.

В данной работе будет рассмотрен морфологический анализ имен существительных. Как уже отмечалось, в крымскотатарском языке нет приставок и окончаний и, кроме этого, существительные в именительном падеже единственного числа являются корнем слова.

Образование словоформ происходит по следующей схеме:

1. корень, т.е. само слово
2. словообразующие аффиксы
3. формообразующие аффиксы¹:
 - число
 - уменьшительно-ласкательные
 - обладания/лишения
4. словоизменяющие аффиксы²:
 - принадлежности
 - сказуемости
 - падежности

Формообразующие аффиксы [3]:

- *Число.* Для всех частей речи в крымскотатарском языке множественное число характеризуется наличием аффиксов *-лар, -лер* (правило сингармонизма [2]).

Например, балалар – дети, еллер – ветра.

- *Уменьшительно-ласкательные.* Для существительных в крымскотатарском языке уменьшительно-ласкательная категория характеризуется наличием аффиксов *-чыкџ, -чик* (правило сингармонизма).

Например, балачыкџ – ребеночек, ельчик – ветерочек.

- *обладания/лишения.* Для существительных в крымскотатарском языке категория обладания (лишения) характеризуется наличием аффиксов *-лы (-сыз), -ли (-сиз)* (правило сингармонизма).

Например, балалы (баласыз), елли (ельсиз).

Словоизменяющие аффиксы [3]:

¹Формообразующие аффиксы могут использоваться как по отдельности, так и вместе, причем порядок зависит от контекста

²В данном случае порядок важен и аффиксы принадлежности с аффиксами сказуемости вместе не используются.

- *Принадлежности.* В зависимости от числа и лица используются различные аффиксы:

Число	Лицо	Аффиксы принадлежности
Ед.	1	-м, -ым, -им, -ум, -юм
	2	-нџ, -ынџ, -инџ, -унџ, -юнџ
	3	-сы, -си, -и, -у, -ю
Множ.	1	-мыз, -миз, -ымыз, -имиз, -умыз, -юмыз, (-лар)-мыз, (-лер)-имиз
	2	-нџыз, -нџиз, -ынџыз, -инџиз, -унџыз, -юнџиз (-лар)-нџыз, (-лер)-инџиз
	3	-сы, -си, -ы, -у, -ю, (-лар)-ы, (-лер)-и

- *Сказуемости.* В зависимости от числа и лица используются различные аффиксы:

Число	Лицо	Аффиксы сказуемости
Ед.	1	-м, -ым, -им
	2	-сынџ, -синџ
	3	-дыр, -дир, -тыр, -тир
Множ.	1	-мыз, -миз, (-лар)-мыз, (-лер)-миз
	2	-сыз (-сынџыз), -сиз (-синџиз)
	3	-(-лар)-дыр, -(-лер)-дир, -тыр, -тир

- *Падежности.*

Падеж	Аффиксы
Именительный	- -
Родительный	-нынџ, -нинџ
Дательный	-гџа, -кџа, -ге, -ке
Винительный	-ны, -ни
Местный	-да, -де, -та, -те
Исходный	-дан, -ден, -тан, -тен

Учитывая агглютинативность крымскотатарского языка, представляется интересным сформировать иерархию классов аффиксов (суффиксов), которая имеет важное значение для морфологического анализа слов в предложении. Хотя и не всегда, основываясь только на информации об аффиксах и их иерархии, можно однозначно провести морфологический анализ слов, однако, можно попытаться сузить область неопределенности [4] и применить имеющуюся информацию о морфологии слова для дальнейшего сужения на последующих этапах обработки предложений.

Ниже приведем один из вариантов кода на языке Visual Prolog выше описанных идей для морфологического анализатора имен существительных крымскотатарского языка:

```

class facts
  affixType : (string TypeName, string ParentTypeName, string Caption, integer Priority).
  affix      : (string Title, string Morpho).
  word       : (string Word).

clauses
  % Описание иерархии типов аффиксов
  affixType("aftKelishlik", "", "Kelishlik", 0). % аффикса падежности
  affixType("aftBashKelishi", "aftKelishlik", "Bash Kelishi", 1). % именительный падеж
  affixType("aftSaiplikKelishi", "aftKelishlik", "Saiplik Kelishi", 1). % родительный падеж
  affixType("aftDogrultuKelishi", "aftKelishlik", "Dogrultu Kelishi", 1). % дательный падеж
  affixType("aftTushumKelishi", "aftKelishlik", "Tushum Kelishi", 1). % винительный падеж
  affixType("aftErKelishi", "aftKelishlik", "Er Kelishi", 1). % местный падеж
  affixType("aftChikishKelishi", "aftKelishlik", "Chikish Kelishi", 1). % исходный падеж

  affixType("aftMulkiyet", "", "Mulkiyet", 0). % аффикс принадлежности
  affixType("aftMulkiyetTeklik", "aftMulkiyet", "Teklik", 1). % аффикс принадлежности ед. ч.
  affixType("aftMulkiyetTeklik1Shahis", "aftMulkiyetTeklik", "1 Shahis", 2). % 1-е лицо.
  affixType("aftMulkiyetTeklik2Shahis", "aftMulkiyetTeklik", "2 Shahis", 2). % 2-е лицо.
  affixType("aftMulkiyetTeklik3Shahis", "aftMulkiyetTeklik", "3 Shahis", 2). % 3-е лицо.
  affixType("aftMulkiyetChokluk", "aftMulkiyet", "Chokluk", 1). % аффикс принадлежности мн. ч.
  affixType("aftMulkiyetChokluk1Shahis", "aftMulkiyetChokluk", "1 Shahis", 2). % 1-е лицо.
  affixType("aftMulkiyetChokluk2Shahis", "aftMulkiyetChokluk", "2 Shahis", 2). % 2-е лицо.
  affixType("aftMulkiyetChokluk3Shahis", "aftMulkiyetChokluk", "3 Shahis", 2). % 3-е лицо.

  ...

  % Обучение: какие аффиксы у каких типов
  affix("", "aftBashKelishi").
  affix("нынъ", "aftSaiplikKelishi").
  affix("нинъ", "aftSaiplikKelishi").
  ...
  affix("тир", "aftHaberlikChokluk3Shahis").

  % Словарь
  word("азбар").
  ...
  word("бахыт").

class predicates
  divider : (string Word, list List) nondeterm anyflow.
  root     : (string Root, list AffixList) nondeterm anyflow.
  analysis : (string Name, list Result) nondeterm anyflow.

clauses
  % разбиение входного значения (слова) на всевозможные комбинации аффиксов и корень слова
  divider(_, []).
  divider(Word, [Affix|List]) :-
    ListAffix = toCharList(Affix), % преобразуем строку в список символов
    ListWord  = toCharList(Word),
    conc(ListWordPart, ListAffix, ListWord),
    list(ListWordPart, ListWordPart1),
    PartWord = concatList(ListWordPart1) % преобразуем список в строку string
    divider(PartWord, List).
  divider(Word, [Word|List]) :-
    divider(Word, List), !.

  % выделение корня из полученной комбинации вводимого слова
  root(Root, List) :-
    lastElement(List, Root).

```

```
% анализ списка аффиксов, определение морфологических признаков
analysis(_, []).
analysis(Name, [R1|Result]) :-
    affixType(Name, Name1, R1, Pri),
    analysis(Name1, Result).

% перебор всех аффиксов базы и выделение корректных,
% используя analysis получаем конечное морфологическое представление слова
conclusion(_, []).
conclusion([L|List], [Res|Result]) :-
    affix(L, Name),
    analysis(Name, Res),
    conclusion(List, Result).
```

ЗАКЛЮЧЕНИЕ

В настоящее время не существует теоретических разработок и прикладных проектов по построению информационных систем автоматического понимания крымскотатарского языка. В данной статье делается попытка начать построения математико-информационной модели таких систем. Первым этапом является морфологический анализ, который является частью лингвистического процессора, и которому (МА), в основном, посвящена данная статья.

В работе рассмотрены основные особенности морфологии крымскотатарского языка и сделана попытка их формализации на основе *иерархии аффиксов* с целью построения автоматизированной системы морфологического анализа. В среде логического программирования Visual Prolog была описана иерархия основных типов аффиксов и разработаны предикаты для выделения морфологической основы слов крымскотатарского языка.

СПИСОК ЛИТЕРАТУРЫ

1. *Евдокимова И.С.* Естественнo-языковые системы: курс лекций. – Улан-Удэ: Изд-во ВСГТУ, 2006. – 92 с.
2. *Меметов А.М.* Земаневий кырымтатар тили. – Симферополь, 2006. – 320 с.
3. *Сейдаметова Н.Д.* Сопоставительная грамматика. – Симферополь, 2007.
4. *Анафиев А.С.* Теория шаблонов в задачах обучения по прецедентам и выбора моделей. Диссертация на соискание ученой степени кандидата физико-математических наук. – 2007. – 135 с.

Статья поступила в редакцию 01.06.2009