

Реализация контекстных соответствий $V:u$, $V:U$, $V:w$, $Y:I$ и $Y:0$ в файле фонологических правил

Д.Ш.Сулейманов, Р.А.Гильмуллин

Академия наук Татарстана, Казанский государственный
университет

selet@bancorp.ru

Аннотация

В статье описываются фонологические правила представления соответствия лексического символа V татарского языка поверхностным символам u , U и w и лексического символа Y поверхностным символам I и 0 при генерации словоформы. Лексический символ V , в отличие от большинства лексических символов, представленных в алфавите файла правил, участвует в формировании только одной аффиксальной морфемы, которая после прибавления к глаголу в форме 2 л. ед. ч. образует номинативную форму. Поверхностные символы u , U и w являются алломорфами соответствующей аффиксальной морфемы. Другой лексический символ Y , также рассматриваемый в данной статье, используется в качестве символа, позволяющего описывать некоторые исключительные ситуации.

1. Введение

Татарский двухуровневый морфологический анализатор - это программа, которая осуществляет разложение произвольных словоформ на морфемы и генерацию словоформ на основе фонологических и морфотактических правил [1]. Анализатор построен на основе программного инструментария РС-KIMMO, использующего лингвистические описания фонологии и морфологии естественного языка (ЕЯ) для обработки словоформ [2]. Используя программный модуль - генератор, РС-KIMMO переводит заданную лексическую форму, составленную из морфем, в поверхностную, а используя программу распознавания, заданную поверхностную форму (т.е. нормальную произвольную словоформу) переводит в лексическую (Рис. 1). Поверхностная форма - это запись словоформы согласно орфографическим правилам, т.е. формирование словоформы из основы и алломорфов. Лексическая форма словоформы - это обобщенная запись, которая выражается последовательностью морфем.

В татарском языке, как и в других тюркских языках с агглютинативной морфологией, при обработке естественно-языковых текстов весьма важным является морфологический анализ, т.е. разложение словоформы на составляющие ее морфемы и генерация словоформы из морфем по правилам фонологии.

Данная статья является третьей из серии запланированных, в которых будут детально описываться правила генерации словоформ из составляющих их морфем, и посвящена раскрытию следующих пяти правил файла фонологических правил относительно соответствий $V:u$, $V:U$, $V:w$, $Y:I$ и $Y:O$. Правила описываются в том порядке, как они представлены в файле фонологических правил. Файл фонологических правил татарского языка состоит из 42 правил, 11 из которых реализуют ситуации, вызванные нарушениями морфологии языка [3]. Описание *Правил (1-7)* приведено в работах [4] и [5]. В этой статье описаны следующие *Правила (8-10)* представления аффикса номинизации, а также *Правила (11) и (12)*, описывающие исключительную ситуацию, связанную с нарушением регулярности.

Рис.1. Структурно-функциональная схема двухуровневого анализатора

Рис.1 отражает структурно-функциональную схему двухуровневого морфологического анализатора. Генератор, используя файл фонологических правил, лексическую запись *bakCa+KE* переводит в поверхностную - *bakCaga*. Распознаватель, используя оба файла - файл фонологических и файл морфотактических правил, словоформу *bakCagar* раскладывает по составляющим и соответствующим им содержательным описаниям: Исем(*bakCa*)+[падеж.афф.(-KE)].

2. Структура правила

Фонологическое правило в РС-KIMMO имеет следующее обобщенное представление:

ПРАВИЛО $V:w < = > ЛК_ПК$, (П)

состоящее из трех частей, где

1) $V:w$ – соответствие лексического и поверхностного символов,

2) $< = >$ - оператор, означающий, что соответствие проявляется всегда и только в этом контексте,

3) $ЛК_ПК$ – выражение, означающее левосторонний и правосторонний контексты (ЛК и ПК, соответственно). Знак подчеркивание "_" обозначает позицию, в котором проявляется соответствие $V:w$, т.е. V заменяется на w в зависимости от контекста.

Кроме приведенного здесь оператора, в РС-КИММО используется еще 3 следующих типа операторов, определяющих взаимосвязь соответствия с контекстом:

$= >$ - означает, что соответствие проявляется только в этом контексте, но не всегда;

$< =$ - означает, что соответствие проявляется всегда в этом контексте, но не только;

$/ < =$ - означает, что соответствие никогда не проявляется в описанном контексте.

Подробное описание структуры фонологических правил приводится в [2].

3. Описание пяти контекстных правил из файла фонологических правил татарского языка.

A? Интерпретация фонологических правил, описывающих соответствия $V:u$, $V:U$
и $V:w$

Следующие три правила описывают в файле правил соответствия $V:u$, $V:U$ и $V:w$:

(8) RULE $V:u \Rightarrow @:BACKV (+:0) (H:0) (CS) +:0_;$

(9) RULE $V:U \Rightarrow @:FRONTV (+:0) (H:0) (CS) +:0_;$

(10) RULE $V:w \Leftrightarrow @:VOWEL +:0 _+:0 C:0 [H:e|H:ы];$

Далее рассмотрим каждый из контекстов, в которых наблюдается соответствия, описываемые правилами (8-10):

(a) $@:BACKV (+:0) (H:0) (CS) +:0_;$

(б) $@:FRONTV (+:0) (H:0) (CS) +:0_;$

(с) $@:VOWEL +:0 _+:0 C:0 [H:e|H:л];$

Контекст (а) Правила (8) утверждает, что лексический символ V соответствует поверхностному символу u если ему предшествует символ $+$, который соответствует символу 0 ; далее может быть любой символ из множества согласных букв CS; далее может идти лексический символ H , на поверхностном уровне соответствующий символу 0 , далее снова символ $+$; далее любой символ из множества заднего ряда гласных BACKV.

Приведем примеры с контекстом (а). Пусть на вход генератора подаются следующие последовательности морфем:

$at+V$ 'стреляй' + [аффикс номинизации]

$kara+Hl+V$ 'смотри' + [аффикс взаимного залога] +

[аффикс номинизации]

В процессе генерации по файлу фонологических правил для первого примера устанавливаются промежуточные соответствия $a:a$, $t:t$, $+:0$, для которых нет специальных правил соответствия; далее работает Правило (8), левый контекст для которого удовлетворяет контексту (а), также для второго примера, сначала устанавливаются промежуточные соответствия $a:a$, $t:t$, $+:0$; далее устанавливается соответствие $H:0$ Правилom (6); далее работает Правило (8), левый контекст для которого удовлетворяет контексту (а) и, следовательно, Правило (8) для выше приведенных примеров устанавливает соответствие $V:u$

(Л.у.): $at+V$

(П.у.): $at0u$

(Л.у.): $kara+Hl+V$

(П.у.): $kara00l0u$

Генератор выдаст следующие поверхностные формы записи:

atu 'стрельба, выстрел'

$karalu$ 'быть осмотренным'

Контекст (б) Правила (9) строится также, как и контекст Правила (8), с той лишь разницей, что множество гласных заднего ряда BACKV заменяется на множество гласных переднего ряда FRONTV:

Примеры:

$kit+V$ 'уйди' + [аффикс номинизации]

$bAr+Hl+V$ 'ударь' + [аффикс взаимного залога] + [аффикс номинизации]

Оба примера лексического соответствия $V:U$ удовлетворяют контексту (б) Правила (9):

Генератор выдаст следующие поверхностные формы записи:

$kitU$ 'уйд'

bAreIU 'столкновение'

Рассмотрим далее контекст (в) Правила (10):

(в) @:VOWEL +:0 _+:0 C:0 [H:e|H:I];

Контекст является лево- и правосторонним. Итак, лексический символ *V* соответствует поверхностному *w*, если ему предшествует лексический символ *+*, соответствующий на поверхностном уровне символу *0*; далее любой символ из множества гласных букв VOWEL. Правый контекст тоже содержит лексический символ *+*, соответствующий поверхностному символу *0*; далее лексический символ *C*, соответствующий поверхностному символу *0*; далее лексический символ *H*, соответствующий либо *e*, либо *I*.

Пример. Пусть на вход генератору подаются следующие лексические формы:

kara+V+HN 'смотри' + [афф. наст. времени] + [афф. притяж. 2-го лица ед. ч.]

fikerlA+V+HN 'рассуждай' + [афф. наст. времени] + [афф. притяж. 2-го лица ед. ч.]

Поверхностные формы будут иметь вид:

karawIN 'твой взгляд'

fikerlAweN 'твое рассуждение'

В? Интерпретация фонологических правил, описывающих соответствие *Y:I*, *Y:0*

Следующие два правила описывают в файле правил соответствия ***Y:I*, *Y:0***:

(11) RULE *Y:I* => BACKV CS_CS (+:0 H:I);

(12) RULE *Y:0* => CS_CS +:0 (@:0)[H:I|H:e];

Правила (11) и (12) лексического соответствия *Y* описывают некоторые исключительные ситуации морфологии татарского языка, возникающие в словоформах при использовании аффиксов притяжательности.

Рассмотрим контекст (б) BACKV CS_CS (+:0 H:I) Правила (11).

Контекст (б) утверждает, что лексический символ *Y* соответствует поверхностному символу *I*, если:

- слева направо ему предшествует любой символ из множества согласных букв CS и далее любой символ из множества заднего ряда гласных BACKV.
- справа налево от символа следует любой символ из множества согласных букв CS, далее символ *+*, соответствующий символу *0* на поверхностном уровне, далее лексический символ *H*, соответствующий поверхностному символу *I*.

Пример:

halYk+Hm 'народ' + [афф. притяж. 1 л. ед. ч.]

Контекст удовлетворяет, условию контекста (б) и генератор выдаст следующую форму:

hallglm 'Мой народ '

Рассмотрим контекст Правила (12):

(в) CS_CS +:0 (@:0)[H:I|H:e];

Контекст (в) утверждает, что лексический символ *Y* соответствует поверхностному символу *0*, если:

1. слева направо ему предшествует любой символ из множества согласных букв CS.
2. справа налево от символа следует любой символ из множества согласных букв CS, далее может быть любой символ, соответствующий поверхностному *0*, далее лексический символ *H*, соответствующий либо *e*, либо *I*.

Пример:

halYk+Hm 'народ ' + [аффт. притяж. 1 л. ед. ч.]

Контекст удовлетворяет, условию контекста (б) и генератор выдаст следующую форму:

halklm ' Мой народ '

Закключение

Итак, в данной статье описаны Правила (8-10) представления аффикса номинизации и Правила (11) и (12), описывающие некоторую исключительную ситуацию, возникающую в связи с нарушениями регулярности морфологии татарского языка. В данном случае по причине неоднозначности орфографии в таких словах как: *hallglm(halklm)* – 'мой народ', *kalktl(kalku)* - 'всплытие', *hollk(hollk)* 'натура', закрепленных, по всей вероятности, при переходе с арабской графики на кириллицу.

Файл фонологических правил, наряду с файлом морфотактических правил [1], является формальной базой татарского двухуровневого морфологического анализатора, выполненного по совместному проекту с лабораторией Билкентского университета в рамках Программы НАТО «Наука за стабильность». В настоящее время двухуровневая модель морфологии татарского языка, включающая описанные в данной статье правила, используется при разработке татарско-турецкого автоматизированного переводчика.

Литература

Suleymanov D.Sh., Gilmullin R.A., Guilmy A.A. «*A Two-level phonological rules of tatar morphology*». KDS-97. Шестая Международная конференция «ЗНАНИЯ - ДИАЛОГ - РЕШЕНИЕ». Сборник научных трудов в двух томах. Ялта - 1997. Том 1. -с.299-305.

Evan L. Antworth. «*PC-KIMMO: A Two-level Processor for Morphological Analysis*». SUMMER INSTITUTE OF LINGUISTICS. Occasional Publications in Academic Computing , 1990.

Сулейманов Д.Ш. Регулярность морфологии татарского языка и типы нарушений в языке. -Казань. Изд-во КГУ. В сб. Трудов «Интеллект. Язык. Компьютер», вып.1,1994. -с. 77-106.

Сулейманов Д.Ш., Гильмуллин Р.А. *Реализация контекстных соответствий А:а, А:ӗ в файле фонологических правил.* // Сборник трудов Математического центра имени Н.И. Лобачевского. Т.4. Компьютерная лингвистика. – Казань: УНИПРЕСС, 1999. – с.127-137.

Гильмуллин Р.А. *Реализация контекстных соответствий Ы:ы, Ы:е и Ы:0 в файле фонологических правил.* // Сборник трудов Математического центра имени Н.И. Лобачевского. Т.4. Компьютерная лингвистика. – Казань: УНИПРЕСС, 1999. – с.51-58.

М Закиев М.З. *Татарская грамматика.* Т2. Морфология. -Казань: Таткнигоиздат, 1992. - 296 с.