

К ПРОБЛЕМЕ ФОНЕМАТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ ОРФОГРАФИЧЕСКОГО ТЕКСТА В СИНТЕЗАТОРЕ ТАТАРСКОЙ РЕЧИ

Т.И.Ибрагимов

Казанский государственный университет

Лингвистическое обеспечение синтезатора татарской речи, точнее преобразователя «Текст - Речь», включает автоматическое транскрибирование входного текста, сегментацию транскрибированного слова на слоги, просодическое оформление речевых тактов (ритмических групп) и просодическое оформление фразы. В статье обсуждаются теоретические вопросы фонематического транскрибирования входного текста.

Фонематическое транскрибирование орфографического текста осуществляется в два этапа. На первом этапе определяется принадлежность входного слова тому или иному языку, на втором - собственно фонематическая транскрипция. Наличие первого этапа - выделения иноязычных слов (неадаптированных заимствований) обусловлено отсутствием научно разработанной орфографии языка. Современная орфография татарского языка представляет эклектику трех языковых систем - собственно татарской, русской и арабской. Данное обстоятельство существенно затрудняет построение фонематического транскрибатора текста и делает невозможным разработку алгоритма автоматического сегментатора слов на слоги. Для того, чтобы правильно транскрибировать, к примеру, слова *корабль* и *кораллы* необходимо знать к каким языкам они принадлежат, точнее, на каком языке они написаны. Далее, как известно, русский язык относится к акцентным языкам, и как таковой, обладает ярко выраженным словесным ударением. Татарский же язык является сингармоническим. Слова в татарском произносятся менее энергично, чем в русском, и без ударения - обязательного выделения какого-либо слога. Если ритмика русской поэтической речи основана на чередовании ударных и безударных логов, то ритмика татарского поэтического текста - на равном количестве слогов в строках и на распределении пауз (цезур) в определенных местах. Указанные различия в ритмической структуре двух поэзий в неявной форме проявляются в русской и татарской спонтанной речи. Это означает наличие определенных различий в слогоделении в этих языках, а

возможно, и в слогеобразовании. Так, если слова *почта* и *республика*, согласно фонетике русского языка, распадаются на слоги как «почта» и «ре-спу-бли-ка», то в татарском языке деление их на слоги будет выглядеть как «поч-та» и «рес-пуб-ли-ка». Согласно системе русского слогаделения, слово *милиция* распадается на слоги «ми-ли-ци-я», а в татарском - «ми-лит-си-я». Из-за расслабленной артикуляции бифокальный звук «ц» переходит в бифонемное сочетание. Необходимость исследования заимствований, их теоретического осмысления диктуется и тем обстоятельством, что эти заимствования неединичны. В лексике современного татарского языка русские заимствования составляют около 25%, арабские заимствования (неадаптированные) - 10%. Тексты деловых бумаг, научного и публицистического стилей на 30% и 20%, соответственно, покрываются заимствованиями из русского языка. Для знакомства с их правописанием и произношением авторы грамматик, орфографических и орфоэпических словарей отправляют читателей к литературе по русскому языку. Поскольку этого нельзя делать по отношению к компьютерам, мы были вынуждены изучить фонологическую структуру татарского языка, привести нормы произнесения в соответствие со звуковой системой языка.

Описание фонетического строя языка предполагает, прежде всего, определение его состава фонем. В фонетике утверждается наличие в татарском языке 12 гласных и 28 согласных. Выделение фонем производится на основании правила: «Если два звука, находящиеся в одинаковых фонетических позициях, позволяют различить смысл слов, то они являются различными фонемами». Это простое и удобное в применении правило не пригодно для тех языков, в которых принято писать и произносить заимствования так, как в оригинале. Связано это с тем, что среди заимствований всегда найдутся слова, контрастирующие со словами заимствующего языка. Так, из оппозиции орфограмм слов *тол* (вдовый) - *тол* (взрывчатое вещество), *сыр* (резьба) - *сыр* (пищевой продукт) и *сет* (раздави) - *сет* (термин из тенниса) лингвисты сделали заключение о наличии в татарском языке двух фонем /o/, двух фонем /ы/ и двух /э/. Таким образом, механическое применение метода квазимонимов (слов, различающихся только одной фонемой) привело не только к увеличению состава гласных, но и к искаженному представлению звуковой системы татарского языка.

«Совокупность фонем данного языка представляет не простой набор разрозненных единиц. Фонемы находятся в определенных отношениях друг к другу, определенным образом связаны между

собой, составляя известную систему» - пишет Л.Р. Зиндер [1, с.74]. Еще более конкретно выражается на этот счет В.А.Богородицкий. В работе «Характеристика звукового состава в Казанском говоре татарского языка в соотношении со звуковой системой русского языка» он пишет: «Так как звуковая система каждого языка представляет не простое собрание звуков, но гармонически связанное целое, то вполне естественно, что те или иные свойства гласных могут быть родственны со свойствами звуков согласных; полагаю, что только что указанная особенность физиологическая и акустическая гласного /а/ (более задняя артикуляция по сравнению с русской /а/ - И.Т.) может находиться в связи с более глубоким произношением татарского /к (q)/ и /г (g)/ [2, с.39]. В нашем случае, т.е. при изучении звуковой системы языка в целях разработки норм компьютерного произнесения фраз, фонемный состав татарского языка определялся на базе учения о языковых единицах Соссюра, Бодуэна де Куртене, Л.В.Щербы и их последователей. Согласно трактовке данных ученых, фонема - минимальная звуковая единица языка, существующая в сознании своих носителей. Предполагается, что фонемы имеют акустический и артикуляторный образы, возникающие в процессах воспроизведения и восприятия словарных и грамматических значений. Иными словами, эти образы находятся в ассоциативной связи с определенными значениями. Так, в работе «Русские гласные» Л.В.Щерба определяет фонему как «...кратчайшее общее фонетическое представление данного языка, способное ассоциироваться со смысловыми представлениями и дифференцировать слова и могущее быть выделяемо в речи без искажения фонетического состава слова...» [3, с. 121].

Опираясь на данное понимание сущности фонемы и учитывая роль речевых звуков в передаче значений, было выделено 9 гласных (/а/, /ə/, /о/, /ə/, /у/, /ү/, /ы/, /е/, /и/) и 23 согласных фонем (/б/, /в/, /г/, /д/, /ж/, /з/, /й/, /к/, /л/, /м/, /н/, /п/, /р/, /с/, /т/, /ф/, /х/, /ш/, /ч/, /ж/, /ц/, /л/, /л/, /л/ в звуковой системе татарского языка.

Очевидно, что состав фонем любого языка должен удовлетворять условиям необходимости и достаточности. Проверка того, в какой степени выделенные фонемы удовлетворяют этим условиям, может быть осуществлена лишь на множестве грамматических морфем. Объясняется это следующими причинами: во-первых, трудно доказать, что квазиомонимы, например, слова [тол (вдовый)] и [тол (взрывчатка)] в фонетическом плане различаются только по оппозиции гласных; во-вторых, не для всех фонем можно подобрать квазиомонимы (например, для фонем /ц/, /ш/, /ж/, /в/ квазиомонимы в

татарском языке не существует) и, в третьих, лексика любого языка представляет открытое множество, т.е. в любой момент она может обогатиться за счет нового слова. В нашем случае анализ выделенных фонем на достаточно большом, но ограниченном множестве грамматических морфем показал, что данный состав фонем удовлетворяет условиям необходимости и достаточности. Фонемы в языке вступают в синтагматические и парадигматические отношения. Одна из проблем, возникающих при исследовании звуковой системы языка, состоит в подборке материала, на базе которого можно было бы изучать эти отношения. Использование письменной речи для этих целей неприемлемо, поскольку она является нормированной, а современная татарская орфография, как уже отмечалось, представляет эклектику трех языковых систем.

Не совсем корректным является выбор устной речи в качестве материала исследования звуковой структуры языка. Как известно, фонетические единицы (речевые звуки) отличаются от фонологических (фонем) многими свойствами, в том числе индивидуальностью, вариативностью и бесконечностью. С учетом сказанного в данной работе изучение структурных ограничений в сочетаниях фонем проводилось на материале (выборке), включающем базовую лексику языка, адаптированные заимствования и грамматические морфемы. В отличие от других компонентов указанные множества составляют социально отработанную часть языка. Базовая лексика татарского языка была составлена на основе списка слов, предложенного Сводешем. Собственно, это переводы 215 слов, выбранных Сводешем как исконные слова любого языка; как нейтральные к предметам материальной культуры и поэтому не испытавшие влияния социально - культурных взаимоотношений наций.

Из анализа выборки следует, что основную роль в формировании звуковой системы татарского языка сыграл фактор удобопроизнесения - стремление минимизировать затраты энергии при произнесении. Об этом, в частности, свидетельствуют такие факты, как отсутствие в базовой лексике слов с нарушением закона сингармонизма, использование вставочных гласных для озвучивания сочетания согласных в заимствованных словах (*стол - өстәл, шляпа - эшләпә, шарф - шарыф*), отсутствие грамматических морфем с гласными /o/ и /ө/ (в конце слова они переходят соответственно в /ы/ и /е/, которые более удобны для произнесения), согласование фонем по признаку звонкость/глухость на стыках морфем, оглушение

согласных в конце слова (*пуд - пот, утюг - утук, очередь - чират, кладовая - келет*) и т.д.

Формированию звуковой системы татарского языка под действием фактора удобопроизнесения способствовали, безусловно, особенности его построения. А именно:

а) Отсутствие словесного ударения. Говорящему нет необходимости осуществлять сознательный контроль над артикуляцией слова. Слова произносятся в автоматическом режиме. Начальный слог - давление набранной в легкие воздушной массы на голосовые связки (межреберных мышц на легкие) несколько больше - произносится более энергично, а конечный слог, по причине уменьшения этого давления, более расслабленно, чем остальные. Все это находит отражение в характере изменения интенсивности и длительности звуковой картины фонетического слова.

б) Отсутствие в татарском языке префиксов. Морфемы в слове располагаются по степени смысловой значимости - корневая морфема, словообразовательные аффиксы и словоизменяющие аффиксы. Говорящему нет необходимости выделять с помощью модуляции голоса корневую морфему, как, например, в русской речи, чтобы довести до слушателя основное значение слова. В этих условиях произнесение слова осуществляется в автоматическом режиме в том смысле, в каком это выражение трактовалось выше.

в) Фонологическая значимость признака мягкость/твердость (палатальность/ непалатальность) в различении корневых морфем (*бала /ребенок/ - б'а'л'а' /беда/, кол /раб/ - к'о'л'/зола/*). Оппозиция корневых морфем по этому признаку обогатила язык гласными звуками. К пяти гласным, имеющимся во всех языках, в татарском добавились еще четыре: /ə/, /ə/, /ʏ/ и краткое /ы/.

Отмеченные выше особенности звуковой системы и построения языка легли в основу алгоритма автоматической фонематической и фонетической транскрипций входного текста. Были учтены также изменения, происходящие в татарском языке. В частности то, что закон сингармонизма в чистом виде, т.е. таком, в каком он существует в базовой лексике, в современном языке не действует. Анализ речи студентов, получивших среднее образование на татарском языке, показал, что согласование звуков по палатальности при произнесении ими заимствованных слов, в целом, распространяется не на все слово, а только на слоги. Иными словами, в современной татарской речи в словах с определенной последовательностью могут чередоваться палатализованные и непалатализованные слоги, но исключается

сочетание палатализованной и непалатализованной фонем в пределах слога.

Алгоритм фонематической транскрипции из-за его большого объема в тексте статьи не приводится. Отметим лишь отдельные, принципиальные, с точки зрения автора, моменты фонематической транскрипции слов орфографического текста.

Работа алгоритма начинается с проверки входного слова (X) на наличие в нем букв «ә», «ө», «ү», «ж», «ң», «һ». При наличии во входном слове этих символов или одного из них, слово относится к разряду Татаро-арабские слова (Хта), при отсутствии их (указанных букв) входное слово определяется как Условно русское заимствование (Хур).

Слово, определенное как Условно русское заимствование, на втором шаге проверяется на содержание букв «ц», «щ» и «ж». Поскольку фонемы /ц/ и /щ/ не характерны для татарского языка, а слова *жәл, жирәбә, тәже* определились на первом шаге как Татаро-арабские, наличие указанных символов говорит о том, что Хур - русское заимствование, и в соответствии с этим должно быть отнесено к разряду Русское заимствование (Хр).

В третьем шаге слово Хур, не содержащее букв «ц», «щ», «ж», проверяется на наличие в его конечной позиции, а также внутри слова в сочетании с другим согласным таких пар букв, как: «зм», «ст», «кс», «нс», «нк», «кт», «шт», «фт», «сс», «бб», «кл», «тр», «лл», «нг», «мн», «мм», «пп», «бр», «нр», «гс», «др». Слова, содержащие сочетания «рт», «лт», «йт» и «нт» в конечной позиции и в середине в соседстве с другой согласной буквой, проверяются на их присутствие в списке таких слов, как *карт, сырт, йорт, илт, кайт, ант* и т.д. (список неполный). При их отсутствии в этом списке слово Хур определяется как русское заимствование.

В четвертом шаге исследуется присутствие в Хур сочетания «ль». Слова, содержащие данное сочетание, затем проверяются на предмет их присутствия в списке лексем типа *шималь, гыйльми* и т.д. (список неполный). В случае отсутствия слова с «ль» в списке слово Хур причисляется к разряду Русское заимствование.

В пятом шаге Хур проверяется на содержание буквы или букв «о».

При наличии во входном слове буквы «о» алгоритм осуществляет следующие проверки:

а) существует ли во входном слове вторая «о»? Если «да», то Хур русское заимствование, если «нет», то Хур сохраняет свой статус, т.е. остается Условно русским заимствованием и как таковое подвергается следующей дополнительной проверке.

б) занимает ли «о» позицию первой или второй буквы от начала слова? Если «да», то Хур проверяется на наличие в списке слов, содержащих такие слова, как, например, *орган, оптика* и т.д., и при обнаружении в этом списке определяется как русское заимствование. Если «нет», то Хур проверяется на наличие в списке слов, содержащих лексемы типа *башкорт, аккош* и т.д., и при отсутствии в этом списке определяется как русское заимствование.

В шестом шаге входное слово, определенное как Условно русское заимствование, проверяется на присутствие буквы «в». При наличии буквы «в» алгоритм ищет ответа на следующие вопросы:

а) встречается ли «в» в начале слова? В случае положительного ответа Хур проверяется на наличие в списке слов типа *вакыт, вакыл, вәҗдан* и т.д. При отсутствии слова в этом списке проверяемое слово причисляется к русским заимствованиям.

б) встречается ли «в» между двумя гласными? Если «да», то Хур проверяется на присутствие в списке слов типа *тавык, кавын, авыл* и т.д. Отсутствие Хур в этом списке говорит о его иноязычности. Если «нет», т.е. «в» не встречается между двумя гласными, то Хур определяется как русское заимствование.

Таким образом, в результате работы первой части алгоритма (приводится в сокращенном виде) входное слово относится либо к разряду «русские заимствования», либо к разряду «арабские заимствования», либо к разряду адаптированных слов.

Второй этап работы алгоритма транскрибирования орфографического текста - преобразование его в фонемную запись - производился на базе богатого теоретического и фактического материала, изложенного в [2], анализа заимствований и речи носителей татарского языка.

Допустим, что слово входного текста Хур содержит букву «ц». Следовательно, оно относится к разряду Русское заимствование. Алгоритм проверяет, какую позицию в слове буква «ц» занимает. Если «ц» следует в начале либо в конце слова, то она заменяется на «с» (цирк - *сырык*, матрац - *матрас*). В том случае, когда «ц» встречается в середине слова, производится еще одна проверка - следует ли «ц» между двумя гласными? Если «да», то она заменяется на «тс» (милиция - *милитсия*), если «нет», то «ц» заменяется на «с» (концерт, концепсия).

В случае, когда русские заимствования заканчиваются на сочетания «зм», «кс», «нс», «нк», «кт», «фт», «тр», «нш», «нг», «гр», «мн», «пс», «нр», «гс», «др», «рк», «кл» либо предшествуют согласной, и, если за ними не следует «ь», то между согласными вставляется гласная «ы»,

если же за ними следует «ь», то между согласными ставится «е». (октябрь, ноябрь). Сдвоенные согласные «сс», «бб», «лл», «мм», «пп» транскрибируются одной согласной фонемой; в сочетании «ст», когда за ним не следует гласная, «т» опускается (коммунис, маторислар).

Следует отметить, что многие из перечисленных сочетаний в прошлом не встречались и в начальной позиции слова. Однако в настоящее время данные сочетания согласных в позиции абсолютного начала слова все чаще произносятся без вставочной гласной (студент, трибуна, бригадир). Связано это с внедрением в послереволюционные годы русского языка во все сферы деятельности и принятием в татарском языкознании орфоэпических и орфографических правил произнесения и написания заимствований, как в оригинале. Данное обстоятельство нашло отражение в фонематической транскрипции текста.

В отношении транскрипции гласных В.А.Богородицкий отмечает, что «Из ударяемых гласных значительные перемены коснулись русских гласных /o/ и /e/, из которых первый в татарском народном языке заменился через /y/, а второй через /и/» [2, с.209]. Далее, «В области неударенного вокализма прежде всего обратим внимание на передачу предупреденного орфографического «о» ... в виде /a/ ... Предупреденный гласный /e/... субституируется в народном татарском произношении через /и/» [2, с.211].

Отмеченные В.А.Богородицким особенности характерны и для современного народного татарского произношения, а, следовательно, являются эталоном и при синтезе речи. В согласии с этим слова «солдат», «лом», «метр», «телефон», «пароход» были транскрибированы, соответственно, как «салдат», «лум», «митыр», «тилифун», «парахут».

Анализ адаптированных - освоенных языком русских заимствований - показывает, что русская фонема /к/ близка по артикуляции и звучанию палатализованному варианту татарской фонемы /к/, т.е. звуку [к']. Слова *капуста*, *корзина*, *комедия*, *книга*, *решётка* и т.д. освоены татарским языком в форме *кәбестә*, *кәрзин*, *кәмит*, *кәнәгә*, *рәшәткә*.

Учитывая данное обстоятельство и то, что сингармонизм в современном татарском языке распространяется не на все слово, слоги с буквой «к» в русских заимствованиях транскрибировались палатализованными вариантами гласных и согласных фонем. Например, командир - кәмандир, компьютер - кәмпютер, закон - закүн, колхоз - кәлхуз, календарь - кәлиндәр и т.д.

В случае преобразования арабских заимствований слоги, содержащие твердые /q/ и /g/, целиком транскрибировались непалатализованными фонемами. Так, слова *мәгълумат*, *тәкъдим*, *тәгълимат*, *нәкъ*, *мәгрифәт* были транскрибированы в форме *маглумәт*, *такдим*, *таглимәт*, *нак*, *магрифәт*.

Анализ заимствований выявил предрасположенность татарского произношения к палатализации. Так, противительный союз «а» в татарском употребляется в форме «ә», *цыган* - *чегән*, *церковь* - *чиркәү*, *бочка* *печкә*, *так* и - *тәки* и т.д. Исходя из данного явления, при преобразовании арабских заимствований после палатализованного слога все остальные слоги слова (если они не содержали твердые /q/ и /g/) транскрибировались палатализованными фонемами. Например, *әдәбиятчы* - *әдәбийәтче*, *мәгълумат* - *маглумәт*, *чәйханә* - *чәйхәнә*, *әмми* - *әммә*, *хәтта* - *хәттә*.

Транскрибирование собственно татарских и адаптированных слов сводится, в основном, к фонемной записи лексем, содержащих сочетания «ау», «әү», букв «я», «ю», «ё», «в», а также «ь» и «ъ». По сравнению с транскрипцией заимствований решение этих задач не представляет больших трудностей.

ЛИТЕРАТУРА

1. Л.Р.Зиндер. *Общая фонетика*. - М.: Высшая школа, 1979.
2. В.А.Богородицкий. *Введение в татарское языкознание*. - Казань: Татгосиздат, 1953.
3. Л.В.Щерба. *Языковая система и речевая деятельность*. - Ленинград: 1974.