Т.И.Ибрагимов
Казанский (Приволжский) федеральный универститет (Россия)
М.Р.Сайхунов
Институт информатики АН РТ

ПИСЬМЕННЫЙ КОРПУС ТАТАРСКОГО ЯЗЫКА С ОЗВУЧЕНИЕМ ВИЗУАЛИЗИРОВАННЫХ ПРЕДЛОЖЕНИЙ КАК ИНСТРУМЕНТ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ

Статья посвящена изучению возможностей применения письменного корпуса татарского языка в лингвистических исследованиях.

Ключевые слова: татарский язык, письменный корпус, транскриптор

В 2010 году в Лаборатории экспериментальной фонетики и прикладной лингвистики были начаты работы по созданию Письменного корпуса татарского языка. К середине 2012 года в сети Internet появилась его первая версия. Объем этой версии составлял 45 млн. слов. Тексты, входящие в корпус, не подвергались аннотированию и стилистической дифференциации, за исключением разметки границ словоформ и предложений. Корпус предназначался, в основном, для решения задач статистической текстологии, машинного перевода, а также для создания интерфейсов систем речевого общения человека с компьютером. Поисковая система выполняла следующие функции:

- находила заданную пользователем словоформу, а также предшествующие и следующие за ней слова (левый и правый контексты данной словоформы, с которыми она может встречаться),

- определяла количество употреблений как заданной словоформы, так и слов, входящих в левый и правый контексты,

- в качестве примера предъявляла исследователю 50 предложений из корпуса, в которых использовалась заданная словоформа.

В последующие годы работы по увеличению объема и обогащению функциональных возможностей Письменного корпуса продолжались. К настоящему времени разработана новая версия Письменного корпуса татарского языка, позволяющая озвучивать визуализированные предложения [http://corpus.tatfolk.ru]. Объем данной версии Письменного корпуса достигает 116 млн. слов или 10 млн. предложений. Поисковая система Корпуса получила существенное расширение. Новая версия позволяет проводить морфологическую лемматизацию словоформы, многократно увеличить число примеров-предложений с участием заданной словоформы (путем повторения операции вывода на экран множества в 50 предложений).

Существенным преимуществом поисковой системы является возможность более точного определения значения заданной словоформы. Поясним это на примере.

Допустим, что переводчик (им может быть система машинного перевода) желает уточнить, в каком смысле используется в тексте словоформа *басмасын.* Она многозначна и может выступить как в значении глагола «пусть не наступит», так и в значении существительных «ступенька» и «издание». Анализ ближайшего окружения словоформы *басмасын* обнаруживает, что в данном тексте ей предшествует слово *яңа (‘новый’)*, а следует за ней слово *узган (‘прошлый’)*, т.е словоформа встречается в сочетании *яңа басмасын узган.* Полученное знание исключает возможность выполнения словоформой *басмасын* функции глагола. Однако сохраняется неопределенность - употребляется ли она в значении «нового издания» или в значении «новой ступеньки».

Новая версия письменного корпуса позволяет некоторое расширение контекстов словоформы, т.е. возможность увидеть - какое слово предшествует

словосочетанию *яңа басмасын узган*, а также - какое слово следует за данным словосочетанием. Для этого достаточно выделить слово *яңа* или слово *узган*. Если обнаружится, что перед словом *яңа* употребляется в той или иной форме слово *китап,* или *журнал*, или *гәзит* (*газета*), то можно считать, что словоформа *басмасын* употребляется в тексте в значении '*издание'*.

Если учесть, что лексическая полисемия в языках является основным препятствием для широкого употребления систем машинного перевода текстов и синхроного перевода речей ораторов, то дополнительные функциональные возможности новой версии Письменного корпуса важны в плане решения многих практических задач.

Новая версия Письменного корпуса татарского языка позволяет одновременно видеть и слышать предложения татарского языка.

Лингвистика последних лет преимущественно работает с письменной формой языка. Письменное общение в некоторой степени выглядит искусственным, оно исключает обстановку, ситуацию и реакцию слушателя. Оно лишено эмоциональной и просодической составляющих, оторвано от самой субстанции языка – членораздельных звуков.

В Письменном корпусе татарского языка с озвучением визуализированных предложений преобразование письменного текста в фонематическую запись производится на основе учения о фонеме и фонетической системе татарского языка одного из основателей Казанской лингвистической школы В.А.Богородицкого.

В учении В.А. Богородицкого о звуковой системе татарского языка важное место занимает понятие «народного говора», т.е. речь ведется о таких единицах языка и формах их произношения, которые отработаны общественностью. На основе исследования татарской речи и наблюдений над произношением своих учеников русских заимствований ученый делает следующие заключения:

1. Вокализим татарского языка содержит 9 полноценных /а/, /о/, /у/, / ы/, /ә/, /и/, /ө/, /ү/ и одну (/ый/) неполноценную (позиционно ограниченную) гласных фонем. Согласно В.А. Богородицкому, /ый/ встречается только в конце слова, заканчивающегося на гласную /ы/. Следует сказать, что при полном произношении слов фонема /й/ в составе конечной /ый/ звучит так же отчетливо, как и в начальной позиции слова. С учетом сказанного в принятой в Корпусе звуковой системе татарского языка значатся 9 гласных фонем.

2. Консонантизм народного говора включает согласные фонемы /б/, /в/, /g/, /д/, /ж/, /з/, /й/, /q/, /л/, /м/, /н/, /п/, /р/, /с/, /т/, /ф/, /х/, /ч/, /ш/, /w/, /ң/, /ж/, /ь/, /к/, /г/. При создании системы озвучения Письменного корпуса встречающиеся в русских заимствованиях твердые согласные /к/ и /г/ были заменены на мягкие /к’/ и /г’/, как это происходило в народном говоре (*капуста – кәбестә, газета – гәзит*)

3. Татарскому языку несвойственно стечение согласных в пределах слога. Следовательно, не существуют в татарском языке и аффрикаты /ц/ и /щ/. В Корпусе при озвучивании слов /ц/ заменяется на /с/ (*матрац – матрас*) или на /тс/ (*полиция – палитсия*), а /щ/ - на однофокусное /ч/ (*плащ – плач*).

Анализ освоенных языком русских заимствований показал, что непалатализованные русские фонемы /к/ и /г/ близки по артикуляции и звучанию палатализованным вариантам татарских фонем /к/ и /г/, соответственно (*кәрзин, рәшәткә, кәмит, кәрниз, гер, мөгәрич, гәрмун*).

Большое значение для татарской фонетики имеет учение В.А. Богородицкого о просодии слова в татарском языке. При построении автоматического транскриптора авторы настоящей работы руководствовались следующими положениями классика: «Из ударяемых гласных значительные перемены коснулись русских гласных /о/ и /е/, из которых первый в татарском народном языке заменился через /у/, а второй – через /и/ [Богородицкий 1953: 209]. Далее, “В области неударяемого вокализма прежде всего обратим

внимание на передачу предударенного орфографического /о/ ⟨...⟩ в виде /а/ ⟨...⟩ Предударный гласный /е/ ⟨...⟩ субституируется в народном татарском произношении через /и/» [Богородицкий 1953: 211].

Следует отметить, что приведенные примеры произношения татарами русских заимствований соответствуют теории языковой эволюции Бодуэна де Куртенэ [Бодуэн де Куртенэ 1963, 2]. Основоположник Казанской лингвистической школы И.А. Бодуэна де Куртенэ считал, что одним из основных факторов, лежащих в основе развития языка, является стремление к удобству или экономии энергии. Согласно Бодуэну де Куртенэ, данная тенденция в языках проявляется в перемещении артикуляции в ротовую полость, а также гласных нижнего подъема вверх. Данное явление находит отражение и в артикуляции татарами русских заимствований, а также в ослаблении лабиализации к концу слова в словах типа *тормышым (моя жизнь)* и *бөреле (с почками)*.

При отсутствии словесного ударения (такого, как в русском языке) и богатом вокализме стремление к удобопроизнесению явилось основным фактором в формировании звуковой системы татарского языка. Так, в основе небной гармонии лежит стремление говорящего сохранить положение массивного артикуляторного органа – корня языка в первоначальном положении, разбиение консонантных сочетаний в начале и конце слова в заимствованиях вставочной гласной (*икс – икес, шкаф – ышкаф*) – стремление озвучивать звуковые сегменты при минимальной затрате энергии.

Литература

Богородицкий В.А. Введение в татарское языкознание всвязи с другими тюркскими языками. /В.А.Богородицкий / Под ред.чл.-кор. АН СССР проф. Н,К,Дмитриева. -2-е изд., испр. и доп. –Казань. Татгосиздат, 1953. -220

Бодуэн де Куртенэ И.А. Язык и языки. Избранные труды по общему языкознанию / И.А. Бодуэн де Куртенэ. – М.: Наука, 1963. – Т.2. – 388 с.

Письменный корпус татарского языка [Электронный ресурс] / Сайхунов М.Р., Ибрагимов Т.И., Салимзянов И.Ф. – Казань, 2012. – Режим доступа: http://corpus.tatfolk.ru