УДК 811.512.1,81’32

КОРПУС ТАТАРСКОГО ЯЗЫКА: КОНЦЕПТУАЛЬНЫЕ И ЛИНГВИСТИЧЕСКИЕ АСПЕКТЫ

© Д.Ш.Сулейманов, Б.Э.Хакимов, Р.А.Гильмуллин

В статье обсуждается концепция корпуса татарского языка, предлагается модель корпуса, рассматриваются вопросы представления лингвистической информации и принципы морфологической разметки татарских текстов. В качестве отдельного аспекта исследуется проблема репрезентативности текстовой коллекции корпуса, предлагается статистический подход к репрезентативности. Вопросы разработки корпуса татарского языка рассматриваются в связи с особенностями языковой системы.

Ключевые слова: татарский язык, тюркские языки, информационные технологии, корпус татарского языка, лингвистические модели, лингвистическая разметка в корпусе, репрезентативность.

Введение

Исследование естественных языков на эмпирически достоверном материале с использованием технологий автоматической обработки языковых данных представляет перспективное междисциплинарное направление в современной науке. Одним из эффективных средств решения многих лингвистических задач являются электронные корпусы языков. Создание подобной системы для татарского языка позволяет получать новые данные о структуре языка, о его лексическом составе и дает ценный материал для дальнейших исследований в области построения лингвистических моделей и реализации технологий автоматической обработки текстов не только на татарском, но и на других естественных языках, в первую очередь, тюркских.

В настоящее время проекты создания общедоступных корпусов тюркских языков особенно актуальны. Тюркская корпусная лингвистика находится на начальной стадии развития, о чем свидетельствует небольшое количество по-настоящему репрезентативных корпусов текстов на тюркских языках. Для татарского языка в данный момент практически не существует каких-либо специализированных лингвистических корпусов. Этим, в первую очередь, и определяется актуальность подобных исследований и разработок не только для татарского языкознания, но и для тюркологии в целом.

Разрабатываемый корпус татарского языка имеет определенную предысторию. Исследования по разработке машинного фонда татарского языка ведутся с начала 90-х годов XX века в Казани научными группами в ряде институтов Академии наук РТ и в Казанском (Приволжском) федеральном университете (подробнее об этом см.: [1; 2]). Основными направлениями традиционно являются разработки формально-алгоритмических компьютерных лингвистических моделей и прикладные исследования в таких сферах, как машинный перевод, поисковые системы, компьютерная лексикография. Концепция Машинного фонда татарского языка, предполагающая создание словарно-грамматического, иллюстративно-текстового и других подфондов, трансформируется и конкретизируется в направлении концепции национального корпуса татарского языка с интегрированными лингвистическими ресурсами.

1. Концептуальная модель корпуса татарского языка

Естественный язык – сложная семиотическая система, которая служит целям коммуникации в человеческом обществе. Существуют различные подходы к моделированию тех или иных явлений языка, который сам по себе является своеобразной информационной моделью мира. Учитывая невозможность универсального формального описания языка, нами был предложен прагматически-ориентированный подход к разработке лингвистических моделей, определяющий минимальный набор средств для решения определенного круга задач, исходя из принципа "контекстной достаточности". Прагматически-ориентированные лингвистические модели классифицируются следующим образом:

1) семиотические модели, обеспечивающие глубинное проникновение в текущий контекст и его трансформацию и использующие помимо лингвистических и экстралингвистические модели, включая "модель контекста" и "модель картины мира";

2) интерактивные модели, обеспечивающие естественно-языковой диалог автоматизированной системы с человеком, включающие коммуникативную и контекстную модели и минимальную лингвистическую информацию;

3) концептуально-формальные модели, обеспечивающие целевую обработку текстов согласно формальным правилам определенного языкового уровня;

4) концептуально-функциональные модели, являющиеся универсальными описаниями значимых единиц определенного уровня или уровней естественного языка [3].

Корпус татарского языка в свете данной классификации представляется как ***интегрированный комплекс концептуально-функциональных моделей различных уровней татарского языка***. Будучи открытой системой, корпус позволяет добавление других лингвистических моделей и автоматизированную разметку на их основе (морфология, синтаксис, семантика и т.д.). Таким образом, в корпусе объединяются морфологическая, синтаксическая, семантическая и др. модели.

Информационная модель электронного корпуса включает морфологическую, синтаксическую, семантическую и другую лингвистическую информацию. Все эти типы информации взаимосвязаны, информация структурирована в форме лингвистических параметров, выраженных при помощи специальных формальных обозначений, обеспечивающих их автоматизированную обработку (разметка). Носителями параметров лингвистической информации являются единицы корпуса.

И наконец, ***модель корпуса представляется как технологическая схема***, которая в общем виде состоит из системы обработки (подготовки) данных для корпуса, собственно корпуса (система баз данных на основе коллекции размеченных текстов) и системы работы с корпусом (управление базой данных, поиск лингвистической информации и др.).

2. Репрезентативность корпуса: подход к решению проблемы

Справедливым является мнение о том, что репрезентативность, т.е. лингвистическая представительность, есть важнейшее свойство любого корпуса и необходимое условие его разработки. Существуют различные стратегии достижения репрезентативности, ориентированные на дискурс, функционирование языка, либо на отображение многообразия языковых явлений, в том числе редких и пассивных языковых единиц [4; 5].

Что касается корпуса татарского языка, то уже на первоначальном этапе коллекция текстов должна обладать определенной репрезентативностью и позволять проводить адекватные исследования. Данная проблема может быть решена в аспекте сбалансированности текстов с точки зрения представленности разнообразных языковых явлений, то есть, чтобы достичь репрезентативности корпуса с ограниченным объемом, необходимо отдавать приоритет текстам, в которых максимально представлены разнообразные, в том числе и нечастотные единицы языка. Для достижения этой цели необходимо производить подробный статистический анализ текстов на этапе включения их в корпус.

2.1. Статистические особенности экспериментальной коллекции татарских текстов

Нами было предпринято статистическое исследование различных квантитативных и лексико-грамматических параметров в экспериментальной коллекции татарских текстов, общий объем которой составляет более 600000 словоупотреблений [6]. Рассмотрим некоторые результаты этого исследования в сфере распределения частей речи.

Таблица 1.

Статистика частей речи: художественная проза (%)

	Сущ.	**Глаг.**	**Прил.**	**Нар.**	**Мест.**	**Служ.**
Т.Галиуллин	39,32	35,68	9,85	1,52	7,4	6,22
Г.Гильманов	35,13	36,71	8,02	2,22	9,94	7,99
А.Еники	35,05	33,75	9,14	2,59	9,56	9,91
Р.Зайдулла	35,6	37,08	7,86	2,33	9,27	7,86
Г.Кутуй	35,51	37,03	8,2	2,58	8,03	8,65
Г.Баширов	34,8	35,76	8,83	2,93	9,58	8,1
Ф.Амирхан	35,17	35,27	8,52	2,86	9,38	8,8
Сводный текст	36,09	36,06	8,57	2,29	9	7,99

Как видно из табл.1, практически все тексты обнаруживают примерно одинаковое распределение частей речи. Сводный текст также дает совпадающие показатели. Эти тексты имеют различный объем, написаны в разное время на протяжении XX столетия. Таким образом, представленные в таблице статистические параметры являются своего рода константами, определяющими "среднюю норму" для данного стиля (в данном случае художественной прозы). В результате появляется возможность измерения степени отклонения статистических характеристик конкретных текстов от этих констант.

Например, с одной стороны, сравнительно высокая частотность существительных, в меньшей степени прилагательных, а с другой стороны – меньший удельный вес наречий, местоимений и служебных частей речи в тексте №1 (рассказы

Т.Галиуллина) могут свидетельствовать о своеобразном употреблении в данном тексте определенных языковых конструкций, а значит, этот текст при включении в корпус усиливает разнообразие и в принципе должен влиять на репрезентативность корпуса.

Таблица 2.

Статистика частей речи: учебная литература (%)

	Сущ.	**Глаг.**	**Прил.**	**Нар.**	**Мест.**	**Служ.**
Текст 1	39,67	32,17	12,43	1,37	6,2	8,16
Текст 2	38,73	34,97	10,99	1,34	5,66	8,31
Текст 3	41,75	39,4	9,15	0,88	3,17	5,65
Сводный текст	39,74	34,81	11,17	1,26	5,34	7,69

В учебных текстах также прослеживаются четкие тенденции, правда, при большем разбросе значений. Последний факт, по нашему мнению, объясняется – субъективно – малым числом текстов данного жанра в коллекции и объективно – неоднородностью самих учебных текстов, обладающих разной степенью "научности" и относящихся к разным предметным областям. Тем не менее, квантитативными маркерами данного жанра могут служить более ярко выраженная номинативность, активность прилагательных и, наоборот, пассивность местоимений по сравнению с художественной прозой.

Таблица 3.

Статистика частей речи: газетные тексты (%)

	Сущ.	**Глаг.**	**Прил.**	**Нар.**	**Мест.**	**Служ.**
Текст 1	40,48	32,21	10,01	1,68	6,75	8,89
Текст 2	45,3	30,26	10,36	1,34	5,44	7,3
Сводный текст	42,47	31,4	10,16	1,54	6,2	8,23

Как следует из таблицы, газетные тексты также обнаруживают своего рода "константы" и достаточно определенные жанрово-стилевые маркеры, позволяющие измерять отклонения конкретных текстов и определять относительную степень их языкового разнообразия.

Таблица 4.

Статистика частей речи: сводные данные по жанрам (%)

	Сущ.	**Глаг.**	**Прил.**	**Нар.**	**Мест.**	**Служ.**
Проза	36,09	36,06	8,57	2,29	9	7,99
Учебные тексты	39,74	34,81	11,17	1,26	5,34	7,69
Газетные тексты	42,47	31,4	10,16	1,54	6,2	8,23
Сводный текст	38,51	35	10,02	1,71	6,91	7,86

Итак, можно полагать, что репрезентативность корпуса текстов ограниченного объема на языке с неоднородным функциональным применением в определенной степени достижима на основе анализа квантитативных и статистических характеристик текстов. Данный метод может быть усовершенствован за счет разработки более точной классификации параметров репрезентативности.

3. Морфологическая разметка в корпусе татарского языка

Принадлежность к агглютинативному типу определяет особенности морфологического строения слова в татарском языке: словоформы формируются путем последовательного присоединения к основе словообразовательных и словоизменительных аффиксов. Каждое грамматическое значение, как правило, выражается отдельным аффиксом, аффиксы регулярны, т.е. способны присоединяться ко всем словам определенной части речи. По этой причине для формального представления морфологии используется анализ аффиксальной цепочки, в некоторых случаях с привлечением словаря основ.

3.1. Модель морфологической разметки татарских текстов

Разработанные к настоящему моменту морфологические модели татарского языка задают определенные стандарты представления татарской агглютинативной морфологии [7; 3]. Автоматизация процесса морфологической разметки текстов в корпусе татарского языка обеспечивается путем создания специализированных программных средств, использующих в своей работе лингвопроцессоры. В частности, в качестве центрального компонента системы автоматизированной морфологической разметки используется адаптированный анализатор на базе двухуровневой модели морфологии татарского языка [8; 9].

Модель морфологии татарского языка базируется на представлении татарской словоформы в качестве определенной последовательности, состоящей из множества аффиксов, и реализована в виде базы морфотактических правил, построенной на основе глагольных и номинативных парадигм и определяющей взаимосвязи между основой и аффиксальными группами языка [8].

Проанализируем возможности автоматизации морфологической разметки татарского текста. Грамматические характеристики татарских частей речи достаточно полно описаны в многочисленных трудах по татарскому языкознанию [10]. Данные характеристики имеют следующее формальное выражение с точки зрения их распознавания и автоматической разметки:

I. Лексико-грамматические (семантические) разряды не имеют явных и однозначных формальных признаков, некоторые из этих разрядов могут иметь непостоянные формальные признаки. Более того, подобная структура произвольной словоформы не указывает однозначно на соответствующий семантический разряд.

II. Собственно морфологические категории в татарском языке имеют явные формальные признаки, которые во многих случаях однозначно характеризуют словоформу.

Следовательно, в татарском языке морфологические характеристики относительно легко распознаются при автоматическом анализе аффиксального состава словоформы. Задача же выявления и автоматической разметки семантических (лексико-грамматических разрядов) различных частей речи не может быть решена с учетом лишь морфологических данных. Исходя из этого, можно полагать, что, по крайней мере, на первом этапе создания корпуса морфологическая разметка может содержать информацию о грамматических категориях, явно выраженных аффиксами.

3.2. Нерегулярные явления в морфологии татарского языка

Рассмотрим проблему так называемых "нарушений" в регулярной морфологии татарского языка и возможности их обработки в процессе автоматической разметки текстов. С точки зрения отношения к системе языка данные нарушения можно подразделить на 2 вида: внешние (несистемные) и внутренние (системные).

1) К внешним нарушениям морфологии татарского языка относятся, в первую очередь, правила морфологического изменения неассимилированных заимствований и случаи, вызванные несовершенством современной татарской орфографии. Типы подобных нарушений описаны в работе [11]. В условиях существующей татарской орфографии и действующих принципов освоения заимствований возникает проблема автоматической обработки этих случаев. Необходимо решить, как обрабатывать заимствования, и найти способы описания возможных закономерностей их изменения.

2) Помимо указанных внешних факторов существуют внутренние языковые особенности, которые являются системными и потенциально автоматически распознаваемыми. Подобные особенности есть следствие универсальности, полифункциональности, экономичности, присущих языковым элементам, что является их системным свойством и подчиняется определенным глубинным закономерностям.

Исходя из особенностей морфологии татарского языка, можно сказать, что трудности при автоматической разметке способны создать полифункциональные и омонимичные аффиксы, так называемые "нулевые" формы, особенности однородных членов (аффикс может присоединяться только к последнему члену группы), особенности изменения отдельных категорий слов (в частности, некоторых местоимений и послеложных слов).

3.3. Стандартизация морфологической информации в корпусе татарского языка

Большое значение для систем автоматической обработки естественного языка имеет разработка концептуальных и технологических стандартов. Морфологический стандарт корпуса татарского языка должен обеспечивать единообразное представление информации, составлять теоретическую основу морфологической аннотации, а также содержать решения, принятые относительно формы представления тех или иных морфологических явлений татарского языка, в том числе тех, интерпретация которых традиционно является дискуссионной.

С точки зрения структуры представления морфологической информации, подобные дискуссионные моменты можно разделить на вопросы, касающиеся частеречной принадлежности лексем, и вопросы, связанные с трактовкой морфологических характеристик словоформ.

В татарской грамматике можно обозначить два основных проблемных случая частеречной характеристики. Во-первых, это вопрос о частеречной принадлежности слов в безаффиксальной форме, выполняющих атрибутивную функцию. Целесообразно, да и более корректно с точки зрения языка было бы классифицировать эти случаи как существительные, не порождая излишнюю омонимию.

Второй "спорный" вопрос представляется более сложным и касается разграничения частей речи (прилагательных и наречий, существительных и прилагательных) во взаимозаменяемых функциях. В принципе допустимы два подхода: рассматривать эти случаи как частеречную омонимию и разрешать ее контекстно, либо присваивать каждой конкретной лексеме однозначную характеристику. Оба подхода имеют как преимущества, так и недостатки и требуют отдельного рассмотрения.

Одним из наиболее сложных является вопрос о падежной системе татарского языка. В последних исследованиях, посвященных данной проблеме, постулируется принципиальная невозможность полного решения этого спорного вопроса в рамках существующей теории [12]. С точки зрения аффиксального состава словоформ, проблема татарских падежей имеет два аспекта:

интерпретация нулевой формы и вопрос о падежном статусе некоторых аффиксов. По нашему мнению, необходимо соблюдать баланс между объективной реальностью языка и традициями грамматической теории. Так, безаффиксальная форма может быть во всех случаях интерпретирована как "основной падеж", аффиксы со спорным статусом должны хотя бы с оговоркой быть включены в падежную парадигму.

Еще одним сложным случаем является категория залога татарского глагола, а именно двоякая природа залоговых аффиксов, которые могут выполнять как сугубо словоизменительную, так и словообразовательную функцию. В результате возникает проблема при лемматизации глаголов с залоговыми аффиксами. В данном случае, очевидно, должна обеспечиваться альтернатива, и в качестве основы можно указывать более чем одну глагольную форму. Такую же альтернативность целесообразно было бы заложить в написании форм инфинитива, так как для многих татарских глаголов наблюдается вариативность в данном вопросе.

Наконец, в татарском языке существует значительное количество полифункциональных аффиксов, часто не учитываемых в традиционных описаниях особенностей словоизменения (например, -лЫ, -сЫз, -чА и др.). Их принадлежность к той или иной категории в татарском языкознании является предметом дискуссии. По нашему мнению, несмотря на то, что такие аффиксы окончательно не включены в татарском языкознании в систему грамматических категорий, их функции должны быть отражены в морфологической разметке корпуса.

3.4. Формат морфологической разметки в корпусе татарского языка

Морфологическая информация о татарской словоформе в корпусе в общем виде состоит из двух основных "полей": 1) частеречная характеристика; 2) совокупность морфологических признаков (параметров).

Учитывая особенности морфотактики в татарском языке, данные параметры можно подразделить, с одной стороны, на сложные и простые, а с другой стороны, на обязательные и факультативные. Сложные параметры могут быть представлены несколькими значениями и, как правило, представляют какую-либо грамматическую категорию (например, система падежных аффиксов). Простые параметры могут быть представлены одним единственным значением (например, вопросительная форма на -мЫ). Что касается обязательности или факультативности параметра, то обязательный параметр должен быть приписан любой словоформе определенной части речи (существительные обязательно стоят в каком-либо падеже, нет "безпадежных" форм существительных). Факультативный параметр, помимо какого-либо конкретного значения, может также принимать отрицательное значение (отсутствие признака – существительные в татарском языке могут и не выражать значение принадлежности).

Заключение

Разработаннная концептуальная модель корпуса татарского языка и модель морфологической разметки татарских текстов позволяет создавать формализованное представление морфологических характеристик произвольного текста на татарском языке в электронном машиночитаемом формате. На первоначальном этапе создания корпуса обоснованным представляется ограничиться в морфологической разметке информацией о грамматических категориях, явно выраженных аффиксами. Эффективность же автоматической разметки можно повысить путем введения дополнительных фонологических и морфотактических правил для морфологического анализатора.

Следующие исследовательские задачи представляются перспективными:

1) разработка в рамках корпуса средств контекстного анализа словоформ;

2) анализ возможностей и разработка правил автоматического определения значений полифункциональных и омонимичных аффиксов, "нулевых" форм, правил автоматического распознавания однородных групп;

3) исследование возможностей создания фонологических и морфотактических правил для специфичных классов лексически, не подчиняющихся общим закономерностям, в первую очередь, заимствований;

4) разработка методов и средств "обучения" лингвопроцессоров и корректировки лингвистических моделей посредством корпусного анализа частотности различных аффиксальных цепочек, анализа случаев присоединения к лексемам одного класса аффиксов из другой парадигмы;

5) разработка синтаксической и семантической моделей корпуса татарского языка, реализация соответствующих систем разметки.

* * * * * * * * * *

1. *Бухараев Р.Г., Сулейманов Д.Ш.* К концепции внедрения татарского языка в компьютерные технологии // Татарский язык и новые информационные технологии. Сер.: Интеллект. Язык. Компьютер. – Казань: Казан. гос. ун-т, 1995. – Вып.2. – С.8-19.
2. *Бухараев Р.Г., Сафиуллина Ф.С., Сулейманов Д.Ш. и др.* К концепции Машинного Фонда Республики Татарстан // Татарский язык и новые информационные

технологии. Сер.: Интеллект. Язык. Компьютер. – Казань: Казан. гос. ун-т, 1995. – Вып.2. – С.20-35.
3. *Сулейманов Д.Ш., Гатиатуллин А.Р.* Структурно-функциональная компьютерная модель татарских морфем. – Казань: Фән, 2003. – 220 с.
4. *Баранов А.Н.* Проблема репрезентативности корпуса данных (на примере политической метафорики) // Тр. Междунар. семинара Диалог-2001 по компьютерной лингвистике и ее приложениям. – Аксаково, 2001. – Т.2. – С.13-15.
5. *Шимкова М.* Репрезентативность корпуса как лингвистическая проблема // Тр. Междунар. конф. MegaLing'2005. – СПб.: Изд-во "Осипов", 2005. – С.130-139.
6. *Гильмуллин Р.А., Невзорова О.А., Хакимов Б.Э.* Корпус татарских текстов: проблема репрезентативности // Тр. Междунар. конф. "Корпусная лингвистика – 2011". 27-29 июня 2011 г., Санкт-Петербург. – СПб.: С.-Петербург. гос. ун-т, 2011. – С.125-130.
7. *Suleymanov D.Sh., Guilmoulline R.A., Guilmoulline A.A.* Tatar phonological rules as a base of two-level morphological analyzer // Proceedings of LP'2000. – Prague: The Karolinum Press, 2000. – P.495-504.
8. *Гильмуллин Р.А.* Разработка файла морфотактических правил для глагольных групп татарского языка // Проблемы сохранения языка и культуры в условиях глобализации: материалы VII Междунар. Симпозиума "Языковые контакты Поволжья" / науч. ред. И.А.Гилязов. – Казань: Казан. гос. ун-т, 2009. – С.222-226.
9. *Хакимов Б.Э., Гильмуллин Р.А.* К разработке системы параметров морфологической разметки для электронного корпуса татарских текстов // Труды Казанской школы-семинара по компьютерной и когнитивной лингвистике TEL-2008. – Казань: Казан. гос. ун-т, 2009. – С.24-29.
10. Татарская грамматика / ред. М.З.Закиев, Ф.А.Ганиев, К.З.Зиннатуллина. – Казань: Татар. книж. изд-во, 1993. – Т.II. Морфология. – 397 с.
11. *Сулейманов Д.Ш.* Регулярность морфологии татарского языка и типы нарушений в языке // Когнитивная и компьютерная лингвистика / науч. ред. Р.Г.Бухараев и др. – Казань: Казан. гос. ун-т, 1994. – С.77-106.
12. *Сулейманов Д.Ш.* К вопросу о числе татарских падежей // Исследования в лингвистике / науч. ред. Р.Г.Бухараев и др. – Казань: Фән, 1996. – С.70-84.

CORPUS OF TATAR: CONCEPTION AND LINGUISTIC ASPECTS

D.Sh.Suleymanov, B.E.Khakimov, R.A.Gilmullin

The conception of the Tatar language corpus is discussed in the paper. The model of the corpus is proposed and the way of representing linguistic information and principles of morphological annotation of Tatar texts are reviewed. As a specific aspect, the problem of representativeness is investigated and specific statistic approach is proposed. Corpus building issues are analyzed on the basis of the language system characteristics.

Key words: Tatar language, Turkic languages, informational technologies, corpus of Tatar, linguistic modeling, linguistic annotation in corpora, representativeness.

* * * * * * * * * *

Сулейманов Джавдет Шевкетович – доктор технических наук, академик АН РТ, заведующий кафедрой математической лингвистики и информационных систем в филологии Института филологии и искусств Казанского (Приволжского) федерального университета, вице-президент, директор НИИ "Прикладная семиотика" АН РТ.

E-mail: dvdt.slt@gmail.com

Хакимов Булат Эрнстович – кандидат филологических наук, доцент кафедры математической лингвистики и информационных систем в филологии Института филологии и искусств Казанского (Приволжского) федерального университета.

E-mail: khakeem@yandex.ru

Гильмуллин Ринат Абрекович – кандидат физико-математических наук, заведующий отделом когнитивных исследований НИИ "Прикладная семиотика" АН РТ.

E-mail: rinatgilmullin@gmail.com

Поступила в редакцию 21.11.2011