УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА.
СЕРИЯ ГУМАНИТАРНЫЕ НАУКИ

2016, Т. 158, кн. 5
С. 1315–1324

ISSN 1815-6126 (Print)
ISSN 2500-2171 (Online)

УДК 811.512.145'322.3

ИССЛЕДОВАНИЕ ЯЗЫКОВОЙ СЕМАНТИКИ С ПОМОЩЬЮ ФОРМАЛИЗАЦИИ ЗАПРОСОВ К КОРПУСНЫМ ДАННЫМ

А.М. Галиева[1], *О.А. Невзорова*[1,2]

[1]*НИИ «Прикладная семиотика» Академии наук Республики Татарстан, г. Казань, 420111, Россия*

[2]*Казанский (Приволжский) федеральный университет, г. Казань, 420008, Россия*

Аннотация

Статья представляет собой первый опыт описания примеров сложных запросов к поисковой системе Татарского национального корпуса «Туган тел», направленных на исследование явлений языковой семантики. Авторы исходят из представления о том, что грамотно сформулированные запросы к корпусу обеспечивают получение сведений, которые позволяют делать выводы о теоретически значимых закономерностях языковой системы. Система грамматических категорий татарского языка и выражающие их аффиксы рассматриваются как ключ к семантике языка. На конкретных примерах показано, что поисковый функционал Татарского национального корпуса даёт возможность извлекать словоупотребления, соответствующие определённым семантическим критериям, из семантически не структурированной корпусной информации. Для построения специальной выборки корпусных данных требуется умение формулировать сложные запросы на специальном языке, разработанном для поиска материалов в корпусе.

Ключевые слова: корпус, татарский язык, поисковые запросы, грамматика, семантика

Введение

Татарский язык обладает сложным синтаксисом и агглютинативной морфологией. В связи с этим мы исходим из приоритетности корпусных исследований как важного источника данных, интерпретация которых является необходимостью для современного лингвистического описания и получения новых теоретических знаний об организации и функционировании языковой системы. Это поднимает проблему построения грамотно сформулированных запросов к корпусу, оптимальных для решения тех задач, которые стоят перед исследователями.

Преимущества использования корпусных технологий в исследовательской и образовательной деятельности очевидны. Корпус как инструмент значительно упрощает извлечение лингвистических данных и их обработку, даёт возможность объединить формальный, квантитативный и эмпирический подходы при изучении языка. Так, корпусные технологии изменили процессы описания языка, подготовки словарей и учебных пособий, методы изучения неродных (иностранных) языков. Лингвистические описания становятся более полными и точными благодаря тому, что основаны на статистической обработке большого

массива данных. Кроме того, корпус позволяет работать с лингвистическими единицами (словоупотреблениями) в том виде, в каком они встречаются в реальных окружениях, а не только с опорой на языковую интуицию исследователя, которая может дать неполные и даже искажённые сведения о языковых фактах. Следовательно, корпусные технологии способствуют снижению субъективности в отборе и анализе языкового материала, а проверяемость результатов подобного исследования обеспечивает его эффективность, достоверность и повторяемость [1, с. 26].

Целесообразность разработки и применения корпусов определяется следующими основными факторами:

1) большой (репрезентативный) объём корпуса гарантирует типичность информации и обеспечивает полноту представления всего спектра языковых явлений;

2) данные разного типа находятся в корпусе в своей естественной контекстной форме, что создаёт возможности для их всестороннего и объективного изучения;

3) однажды созданный и подготовленный массив данных может использоваться многократно, многими исследователями и в различных целях [2, с. 6].

К настоящему времени для татарского языка разработаны и имеются в открытом доступе два основных корпуса:

1) Письменный корпус татарского языка, созданный в Казанском федеральном университете[1] (см. [3]);

2) Татарский национальный корпус «Туган тел», подготовленный в НИИ «Прикладная семиотика» Академии наук Республики Татарстан[2] (см. [4, 5]).

Приходится констатировать, что, несмотря на наличие указанных общедоступных корпусов, их возможности ещё недостаточно освоены лингвистами-тюркологами. Поэтому остро стоят задачи обучения пользователей базовым принципам работы с лингвистическими корпусами и другими электронными ресурсами, построению основных типов запросов к корпусам для поиска материала, а также методам представления результатов своих исследований.

Цель настоящей статьи – показать, как грамотно сформулированный запрос позволяет получить материал для семантических исследований из семантически не структурированного массива корпусных данных.

1. Общая информация о Татарском национальном корпусе «Туган тел»

Татарский национальный корпус «Туган тел» (далее – ТТ) является лингвистическим ресурсом современного литературного татарского языка. Проект выполняется в рамках Государственной программы «Сохранение, изучение и развитие государственных языков Республики Татарстан и других языков в Республике Татарстан на 2014–2020 годы»[3]. Данный корпус адресован широкому кругу пользователей: лингвистам-тюркологам, типологам, преподавателям

[1] http://corpus.tatfolk.ru/
[2] http://corpus.antat.ru/
[3] http://docs.cntd.ru/document/463305579

татарского языка, деятелям культуры, а также всем, кто изучает татарский язык и интересуется татарской культурой.

Объём корпуса на декабрь 2015 г. составил более 80 миллионов словоупотреблений. Он содержит тексты различных жанров и стилей современного татарского языка (художественная литература, публицистика, официальные документы, учебная литература, научные статьи и др.). Каждый документ имеет метаописание: авторы, их пол, выходные данные, даты создания, жанры, части, главы и др. (подробнее см. [5, с. 89–90]).

Тексты, включённые в ТТ, снабжены морфологической разметкой (представлена информация о части речи основы словоформы и наборе её грамматических характеристик), которая выполняется автоматически с применением модуля двухуровневого морфологического анализа татарского языка, реализованного в программном инструментарии PC-KIMMO. Отметим, что от характера и степени разработанности системы разметки во многом зависят пользовательские возможности. В данном случае она ориентирована на представление всех реально существующих грамматических форм слов, не всегда отражаемых в описательных исследованиях по татарской грамматике либо имеющих различные альтернативные трактовки (см. [6]).

Для формального представления татарской агглютинативной морфологии используется модель, в которой словоформа строится путём последовательного присоединения к основе регулярных словообразовательных и словоизменительных аффиксов. Например, имя существительное имеет следующую регулярную морфологическую структуру: <основа> <множественность> <притяжательность> <падежность> <модальность> [5, с. 90–91].

Система грамматической аннотации Татарского национального корпуса основана на Лейпцигских правилах глоссирования [7] с добавлением тегов и терминов для специфических тюркских категорий [6].

Поисковая система данного корпуса позволяет реализовать поиск по следующим параметрам:

- лемме (лексеме);
- словоформе;
- заданному набору морфологических параметров (строка «Опции»).

Запросы к ТТ формируются на специальном логическом языке, в котором действуют следующие синтаксические правила:

1) запрос представляет собой формулу, в которой могут использоваться такие логические функции, как конъюнкция (,), дизъюнкция (|), отрицание (!);

2) сложная логическая функция задаётся при помощи базовых логических операций и скобочных конструкций;

3) операндами логической формулы могут быть аффиксы, грамматические теги, полные или неполные лексические единицы (например, при вводе лексической единицы *китап** получим множество словоформ с корневой морфемой *китап* ‘книга’ и различными аффиксальными цепочками);

4) операнд может задаваться формулой, в которой используется функция «минус», то есть для выражения *китап** *-китаплар* формируется множество словоформ с корневой морфемой *китап*, из которого исключается словоформа *китаплар* ‘книги’.

Таким образом, пользователь, комбинируя различные параметры поиска, может задавать сложные запросы, обусловленные спецификой своего научного исследования.

2. Грамматика и семантика: некоторые аспекты взаимодействия

В имеющихся татарских грамматиках лингвистические категории и аффиксы описываются изолированно, вне связей друг с другом (см., например, [8–10]). Между тем тюркская словоформа представляет собой цепочку последовательно присоединяемых аффиксов, а языковые категории существуют в тесной взаимосвязи. Мы исходим из точки зрения, что языковая форма и её содержание находятся в сложных отношениях. Грамматические категории и значения реализуются только в конкретных словоформах; соответственно, грамматические формы слов можно рассматривать как особые способы интерпретации языковой семантики.

Как указывает С.Д. Кацнельсон, «сложные взаимодействия грамматических форм с единицами плана содержания и общие закономерности распределения элементов поля по формам могут быть вскрыты лишь в результате сопоставления структуры определённого семантико-грамматического поля со всей совокупностью тяготеющих к данному полю единиц плана выражения» [11, с. 146]. По мнению Г.А. Золотовой, «чтобы слово-лексема стало словом-синтаксемой, то есть конститутивной единицей синтаксического строя, оно поднимается на более высокую ступень абстракции, вступая в ряд со словами близкого категориально-семантического значения в соответствующем предложно-падежном оформлении» [12, p. 699–700].

Набор грамматических функций слова, способы его участия в построении конструкций того или иного типа во многом определяются его значением, при этом характер взаимодействия лексики и грамматики неодинаков для разных типов словоформ. Поэтому, осуществляя отбор и компоновку грамматических, лексических и иных параметров запроса к корпусу для поиска материала, мы можем получить определённый доступ к семантической информации, то есть к семантически детерминированным выборкам. Например, противопоставление предметных и признаковых (атрибутивных) значений татарских существительных обусловлено типом конструкций. Так, существительное в атрибутивной функции в грамматически правильно построенном предложении должно стоять непосредственно слева от определяемого слова и быть в основном падеже или иметь ограниченный набор аффиксов (генетив, абессив/мунитатив, локативный атрибутив, номинализатор *-лЫК*[4]):

– *таш* (=номинатив) *йорт* ‘каменный дом’;

– *Казан* (=номинатив) *университеты* ‘Казанский университет’;

– *Габдулланың* (=генетив) *китабы* ‘книга Габдуллы’.

Разные формы слова могут быть обусловлены многообразными аспектами его общего значения, что в значительной мере предопределяет возможности поиска семантических классов языковых единиц исходя из распределения грамматических признаков. Так, слова с пространственным значением будут

[4] Заглавные буквы в аффиксах обозначают фонологически вариативные элементы аффикса, строчные – инвариантные.

стоять преимущественно в падежах, выражающих пространственные отношения (директив, аблатив, локатив), или встречаться в контекстах, содержащих послелоги с пространственным значением.

Иногда для выделения семантических классов достаточно структуры аффиксальной цепочки словоформы. Комбинируя, например, поисковые запросы *имя прилагательное* + *компаратив*, можно получить выборку контекстов, включающих татарские качественные прилагательные – лексико-грамматический класс прилагательных, называющих качественные признаки предмета, которые могут менять свою интенсивность. Такой поисковый запрос выглядит как конъюнкция тегов (ADJ,COMP).

Текущая версия грамматической разметки корпуса содержит элементы словообразования, что также позволяет получить отдельные семантические классы, обусловленные словообразовательной структурой языковой единицы. Простое сочетание параметров поиска даёт контексты, содержащие, например, существительные со значением «действующее лицо», образованные от существительных при помощи аффикса *-чЫ*. Поисковый запрос в таком случае представляет собой конъюнкцию тегов (N,PROF).

Установление статистически значимых зависимостей между значением аффикса, его грамматическими свойствами (в широком смысле) и семантикой слов, а также выявление комбинаций элементов, являющихся носителями смысла, – важнейшие задачи, стоящие на сегодняшний день перед исследователями семантической системы татарского языка.

3. Примеры запросов к корпусу для изучения семантической организации татарского языка

Рассмотрим некоторые типы поисковых запросов, результаты которых позволяют изучить сложные семантические явления татарского языка.

3.1. Поиск нетривиальных форм компаратива. В силу агглютинативной природы тюркских языков в них наблюдается размывание границ между парадигматическими классами при потенциально неограниченном объёме парадигмы, а также нежёсткое распределение лексики по грамматическим классам и частям речи. Наличие общего инвентаря морфем для слов разных частей речи приводит к отсутствию чёткого морфологического деления именных и глагольных форм, когда именные аффиксы присоединяются к глагольным основам, образуя многочисленные гибридные формы. В частности, аффикс сравнительной степени *-рАК* может присоединяться к основам разных типов.

Как известно, чаще всего названный аффикс прибавляется к основе прилагательного. С точки зрения семантической организации татарского языка наиболее интересны случаи его присоединения к основам других частей речи: существительных, осложнённых падежными аффиксами, глаголов и местоимений. Результаты простого запроса для получения форм компаратива (COMP) в основной массе содержат стандартные прилагательные с аффиксом сравнительной степени, которые в данном случае не представляют интереса и должны быть отсечены. Поэтому сузим поисковый запрос, применив, например, конъюнкцию тегов (V,COMP), посредством чего получим выборку глагольных компаративов –

глаголов, обозначающих действие, которое может иметь разную интенсивность и динамические признаки (например, *тырышыбрак* ‘немного стараясь’ от *тырыш-ып* ‘стараясь’).

Из этих результатов легко выделить словоформы определённой аффиксальной структуры. Так, можно сделать запрос на деепричастные формы, присоединяющие аффикс сравнительной степени (например, *суыныбрак* ‘немного остыв’ от *суын-ып* ‘остыв’). Для этого вводим конъюнкцию тегов (COMP,ADVV_ACC). Чтобы получить выборку из причастий с аффиксом компаратива (например, *охшаганрак* ‘немного походящий’ от *охша-ган* ‘похожий, походящий’), необходима формула (COMP,PCP_PS).

Интересными представляются случаи присоединения аффикса *-рАК* к основам существительных: *кояшкарак* ‘туда, где более солнечно’ (от *кояш* ‘солнце’ + ГА + рАК). Результаты простого запроса *существительное* + *компаратив* (N,COMP) будут существенно «загрязнены» стандартными отсубстантивными атрибутивными формами, которые затруднят работу исследователя. Есть два способа от них избавиться. Во-первых, можно убрать из поискового запроса не интересующие нас атрибутивы, образованные от существительных, с помощью такой формулы, как (N,COMP,!(ATTR_MUN|ATTR_ABES)). Она содержит три конъюнктивных члена, при этом третий член представляет собой отрицание дизъюнкции тегов ATTR_MUN и ATTR_ABES. Во-вторых, если исследователь знает, что аффикс компаратива присоединяется только к существительным в форме пространственных падежей: директива, аблатива и локатива, – он может отразить это в своём запросе, используя формулу с конъюнкцией и дизъюнкцией (N,COMP,(DIR|ABL|LOC)).

Следует отметить: поскольку к настоящему времени в корпусе не снята омонимия, в результаты попадают многочисленные наречные образования, омонимичные формам существительных. Однако их можно попытаться устранить, повысив уровень сложности запроса.

Поисковые запросы, подобные представленным выше, позволяют находить и изучать аффиксальные цепочки, не описанные или описанные очень поверхностно в татарских грамматиках, выявляя определённые корреляции между значением слова и его аффиксальной структурой.

3.2. Анализ полифункциональных аффиксов на примере запросов, содержащих аффикс *–лЫК*. Агглютинативная природа татарской морфологии обусловливает однозначность большей части аффиксов. Тем не менее некоторые из них могут функционировать как грамматические и словообразовательные, в зависимости от аффиксальной цепочки словоформы и её контекстуального окружения; в татарских грамматиках такие аффиксы называют полифункциональными [13, с. 356].

Покажем возможности исследования полифункциональных аффиксов татарского языка на примере аффикса *-лЫК*. Поисковая система ТТ позволяет выделять контексты, содержащие образования с этой частью слова, имеющей разную природу. Разберём наиболее типичные случаи.

1. Отадъективные существительные с аффиксом *-лЫК* типа *авырлык* ‘тяжесть’ (от *авыр* ‘тяжёлый’ + лЫК) составляют наиболее продуктивный тип обра-

зования слов с отвлечённым значением признака в татарском языке. Выборку из них можно получить, комбинируя запрос *имя прилагательное + номинализатор -лЫК*, что в логической формуле представлено как конъюнкция тегов (ADJ,NMLZ).

Из полученных данных можно исключить те или иные типы единиц, если это требуется для исследования. В частности, можно устранить слова, образованные при помощи аффиксов атрибутивов *-лЫ* и *-сЫз*. Поисковая формула будет строиться с использованием конъюнкций, дизъюнкции и отрицания: ((ADJ,NMLZ),!(ATTR_MUN|ATR_ABES)). Таким способом мы исключим, например, следующие слова: *җаваплылык* ‘ответственность’ (от *җавап* ‘ответ’ + -лЫ + -лЫК); *җавапсызлык* ‘безответственность’ (от *җавап* ‘ответ’ + сЫз + -лЫК).

Если же, напротив, необходима отдельная выборка слов с аффиксом *-лЫК*, образованных при помощи аффиксов атрибутивов *-лЫ* и *сЫз*, то в поисковой формуле нужно использовать дизъюнкцию и конъюнкцию: ((ATTR_MUN|ATR_ABES),NMLZ).

2. Атрибутивные формы со значением потенциального признака, образованные путём присоединения аффикса *-лЫК* к причастиям будущего времени, например *барырлык* ‘тот, который сможет пойти’, ‘тот, по которому можно пойти’ (от *барыр* ‘пойдёт’ + лЫК), можно получить, если применить логическую формулу, состоящую из конъюнкции трёх тегов (V,PCP_FUT,NMLZ).

3. Выборка из субстантивированных глагольных дериватов при придаточных в сложных предложениях синтетического типа может быть получена с помощью формулы, которая включает набор из четырёх конъюнктивных членов (V,PCP_PS,NMLZ,(DIR|ACC|ABL,LOC)), причём четвёртый член – дизъюнкция трёх тегов.

Таким образом, варьируя параметры поиска, можно получить контексты, содержащие разные группы образований на *-лЫК* для дальнейшего более детального исследования их семантики.

3.3. Возможности изучения грамматики конструкций на примере серии запросов на изафетные словосочетания. Поисковые запросы, нацеленные на получение контекстов с изафетными конструкциями, имеют более сложный вид и требуют описания параметров не одной, а двух рядом стоящих словоформ. Рассмотрим некоторые случаи.

1. Запрос по поиску конструкций типа изафет II предполагает учёт следующих факторов. Первое слово – существительное – должно стоять в основном падеже. В текущей версии корпусной поисковой системы «Туган тел» номинатив задаётся путём исключения косвенных падежей, как в формуле (N,(!GEN,!DIR,!ACC,!ABL,!LOC)) или (N,!(GEN|DIR|ACC|ABL|LOC)). Слово, находящееся непосредственно справа (нужно задать расстояние от 1 до 1), должно иметь аффикс принадлежности третьему лицу, что определяется формулой (POSS_3SG|POSS_3PL).

2. Запрос по поиску конструкций типа изафет III отличается от предыдущего только первой формулой, которая в данном случае выглядит как (N,GEN), поскольку существительное должно иметь форму родительного падежа. Пара-

метры поиска второго слова задаются уже приведённой выше формулой (POSS_3SG|POSS_3PL), при этом расстояние указывается такое же (от 1 до 1).

Наличие изафетных конструкций – одна из специфических черт синтаксиса тюркских языков. Типы изафета различаются не только оформлением, но и характером семантико-синтаксических отношений между компонентами. Семантические классы слов, формирующих изафетные словосочетания, и отношения между компонентами последних в татарском языке описаны довольно поверхностно и требуют дальнейшего более детального изучения на корпусных данных.

Заключение

Итак, план выражения и план содержания в языке тесно связаны: грамматические категории, как правило, реализуются в определённых типах контекстов и в единицах определённых семантических классов. Появление новых лингвистических ресурсов требует разработки новых методов исследований. Возможности поисковой системы Татарского национального корпуса «Туган тел» в известной мере позволяют извлекать данные, соответствующие семантическим критериям, из семантически не структурированных корпусных материалов. Для этого необходимо умение формулировать сложные запросы, направленные на получение лингвистических единиц, обладающих заданными свойствами. Дальнейшие специальные лингвистические исследования, как представляется, должны быть нацелены на поиск корреляций между уровнем формальной организации языка и грамматической и лексической семантикой.

Благодарности. Исследование выполнено при финансовой поддержке РФФИ (проект № 15-07-09214).

Литература

1. *Борискина О.О.* Корпусное исследование языка: мода или необходимость? // Вестн. Воронеж. гос. ун-та. Сер. Лингвистика и межкультурная коммуникация. – 2015. – № 3. – С. 24–27.
2. *Захаров В.П., Богданова С.Ю.* Корпусная лингвистика. – Иркутск; СПб.: Изд-во СПбГУ, 2013. – 147 с.
3. *Ибрагимов Т.И., Сайхунов М.Р.* Письменный корпус татарского языка: идеи, проблемы, решения // Нематериальное культурное наследие тюркских народов как объект сохранения: Сб. материалов Междунар. науч.-практ. конф. (16–19 июля 2014 г.). – URL: http://corpus.tatar/index.php?openinframe=articles.htm. свободный.
4. *Невзорова О.А., Мухамедшин В.Р., Билалов Р.Р.* Корпус-менеджер для тюркских языков: основная функциональность // Тр. Междунар. конф. «Корпусная лингвистика – 2015». – СПб.: Изд-во С.-Петерб. гос. ун-та, 2015. – С. 344–350.
5. *Сулейманов Д.Ш., Невзорова О.А., Галиева А.М., Гатиатуллин А.Р., Гильмуллин Р.А., Хакимов Б.Э.* Размеченный корпус татарского языка «Туган тел»: аспекты реализации // Тр. Казан. шк. по компьютерной и когнитивной лингвистике TEL-2014. – Казань: Фэн, 2014. – С. 88–93.
6. *Галиева А.М., Хакимов Б.Э., Гатиатуллин А.Р.* Метаязык описания структуры татарской словоформы для корпусной грамматической аннотации // Учён. зап. Казан. ун-та. Сер. Гуманит. науки. – 2013. – Т. 155, кн. 5. – С. 287–296.

7. Leipzig Glossing Rules. Conventions for interlinear morpheme-by-morpheme glosses. – URL: https://www.eva.mpg.de/lingua/resources/glossing-rules.php, свободный.
8. Татарская грамматика: в 3 т. – Казань: Тат. кн. изд-во, 1993. – Т. 2: Морфология. – 397 с.
9. Татар грамматикасы: 3 т. – М.; Казан: ИНСАН: Фикер, 2002. – Т. 2: Морфология. – 447 б.
10. *Хисамова Ф.М.* Татар теле морфологиясе: югары уку йортлары өчен д-лек. – Казан: Мәгариф, 2006. – 335 б.
11. *Кацнельсон С.Д.* О грамматической семантике // Кацнельсон С.Д. Общее и типологическое языкознание. – Л.: Наука, 1986. – С. 145–152.
12. *Золотова Г.А.* Взаимодействие лексики и грамматики в семантической структуре предложения // Revue des études slaves. – 1994. – Т. 66, F. 3. – P. 699–707.
13. Татарская грамматика: в 3 т. – Казань: Тат. кн. изд-во, 1993. – Т. 1: Введение. Фонетика. Фонология. – 584 с.

Поступила в редакцию
25.05.16

Галиева Альфия Макаримовна, кандидат философских наук, ведущий научный сотрудник

Научно-исследовательский институт «Прикладная семиотика» АН РТ
ул. Баумана, д. 20, г. Казань, 420111, Россия
E-mail: *amgalieva@gmail.com*

Невзорова Ольга Авенировна, кандидат технических наук, заместитель директора по науке; доцент кафедры информационных систем

Научно-исследовательский институт «Прикладная семиотика» АН РТ
ул. Баумана, д. 20, г. Казань, 420111, Россия

Казанский (Приволжский) федеральный университет
ул. Кремлёвская, д. 18, г. Казань, 420008, Россия
E-mail: *onevzoro@gmail.com*

ISSN 1815-6126 (Print)
ISSN 2500-2171 (Online)

UCHENYE ZAPISKI KAZANSKOGO UNIVERSITETA. SERIYA GUMANITARNYE NAUKI
(Proceedings of Kazan University. Humanities Series)

2016, vol. 158, no. 5, pp. 1315–1324

Study of Linguistic Semantics by Means of Formalisation of Queries to Corpus Data

A.M. Galieva[a*], *O.A. Nevzorova* [a,b**]

[a]*Research Institute of Applied Semiotics, Tatarstan Academy of Sciences, Kazan, 420111 Russia*
[b]*Kazan Federal University, Kazan, 420008 Russia*
E-mail: [*]*amgalieva@gmail.com*, [**]*onevzoro@gmail.com*

Received May 25, 2016

Abstract

The advantages of using linguistic corpus data in education and research are obvious and well covered in specialized literature. This tool considerably simplifies acquisition of linguistic data and their processing.

Two main corpora have been built for the Tatar language by now, each in open access: the Corpus of Written Tatar compiled in Kazan Federal University, (http://search.corpus.tatar/en) and the Tatar

National Corpus (http://corpus.antat.ru/?lang=en) developed by researchers of the Institute of Applied Semiotics, Tatarstan Academy of Sciences, Russia. These corpora are being hourly replenished; the update of textual collections is mainly carried out through the use of media texts, which provides constant flow of fresh linguistic material.

The Tatar language has complicated syntax and intricate agglutinative morphology, and corpus data is a reliable tool for enriching and deepening linguistic descriptions of Tatar. This paper is the first attempt to describe examples of complex queries to the search system of "Tugam Tel" Tatar National Corpus, these queries are aimed at studying complicated phenomena of Tatar linguistic semantics. The authors proceed from the viewpoint that correctly formulated queries to the Corpus provide data allowing to draw conclusions about theoretically relevant laws of the language system. The inventory of grammatical categories of the Tatar language and affixes that express these categories have been considered as a key to language semantics. The authors, by means of particular examples, have shown that search functionality of the Tatar National Corpus enables to extract data meeting certain semantic criteria, from semantically unstructured corpus data. Construction of special samples of corpus data requires an ability to formulate complex queries in a special language, designed for searching data in the corpus.

Keywords: corpus, Tatar language, search query, grammar, semantics

Acknowledgments. The study was supported by the Russian Foundation for Basic Research (project no. 15-07-09214).

***Для цитирования*:** *Галиева А.М., Невзорова О.А.* Исследование языковой семантики с помощью формализации запросов к корпусным данным // Учен. зап. Казан. ун-та. Сер. Гуманит. науки. – 2016. – Т. 158, кн. 5. – С. 1315–1324.

***For citation*:** Galieva A.M., Nevzorova O.A. Study of linguistic semantics by means of formalisation of queries to corpus data. *Uchenye Zapiski Kazanskogo Universiteta. Seriya Gumanitarnye Nauki*, 2016, vol. 158, no. 5, pp. 1315–1324. (In Russian)