

БАЗА СЕМРОНОВ ТАТАРСКОГО ЯЗЫКА КАК ЦИФРОВОЙ ЛИНГВИСТИЧЕСКИЙ РЕСУРС ДЛЯ ОБЪЯСНИТЕЛЬНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Р. А. Гильмуллин^[0000-0002-8520-8921]

Академия наук Республики Татарстан, г. Казань, Россия

rinatgilmullin@gmail.com

Аннотация

Рассмотрена база семронов татарского языка как цифровой лингвистический ресурс для объяснительного искусственного интеллекта. Предложена модель Большой семиотической базы семронов, в которой лексико-грамматические единицы представлены как семроны – потенциальные смыслообразующие единицы с формой, классом, синтаксическими, семантическими и прагматическими характеристиками, валентностями, правилами активации, корпусными подтверждениями, связями с графом знаний и статусами экспертной проверки. Фреймовое представление использовано как способ организации записи семрона: базовый фрейм фиксирует потенциальные возможности единицы, а специализированный фрейм сейрона описывает ее контекстную реализацию после закрытия одной или нескольких валентностей. Показано, что такая база может выполнять функцию ресурсного слоя Семиотической генеративной распознающей модели, обеспечивая переход от морфемной структуры татарского текста к ситуационной модели, верифицируемой цепочке и объяснению. Компактные примеры записей демонстрируют структуру фрейма, типы слотов и фасетов, принципы верификации и возможные форматы представления данных.

Ключевые слова: семрон, база семронов, большая семиотическая база семронов, татарский язык, цифровой лингвистический ресурс, фреймовая модель, валентность, сейрон, граф знаний, корпус, экспертная верификация, объяснительный искусственный интеллект, семиотическая генеративная распознающая модель.

ВВЕДЕНИЕ

Современные системы обработки естественного языка все чаще применяются для работы с малоресурсными языками, однако качество такой обработки ограничивается не только объемом доступных текстов, но и отсутствием специализированных цифровых ресурсов, способных представлять смысловую структуру языка в машинно-читаемой форме. Для татарского языка эта проблема особенно актуальна: он обладает развитой агглютинативной морфологией, высокой ролью аффиксальных цепочек и контекстно обусловленных лексико-грамматических функций.

Существующие цифровые лингвистические ресурсы решают важные, но разные задачи. Словарь фиксирует лексические значения и грамматические сведения. Корпус предоставляет реальные употребления в языке с частотными данными. Морфологический анализатор позволяет выделить основу, аффиксы и грамматические теги. Граф знаний задает связи между объектами, понятиями и отношениями. Однако для объяснимой обработки текста необходимо не только распознать форму, но и показать, какая смысловая функция реализуется в конкретном контекстном окружении, какие условия привели к ее выбору и каким образом этот выбор может быть проверен.

Для агглютинативного языка различие между морфологическим разбором словоформы и определением смысловой функции морфемы в контексте принципиально. Одна и та же аффиксальная морфема может быть представлена в разных поверхностных вариантах, входить в разные цепочки, вступать в различные синтаксические, семантические и прагматические связи. Поэтому цифровой ресурс, ориентированный на объяснимую обработку татарского текста, должен фиксировать не только форму и значение, но и потенциальные валентности, правила их активации, корпусные примеры, связи с графом знаний и статус экспертной проверки.

В настоящей работе предложена модель базы семронов татарского языка как цифрового лингвистического ресурса. Единицей такого ресурса является семрон – потенциальная лексико-грамматическая смыслообразующая единица с формой, синтаксисом, семантикой, прагматикой и открытыми валентностями.

В Большой семиотической базе семронов (БСБС) семрон представлен в виде базового фрейма, содержащего паспорт, слоты, фасеты, ограничения, валентности и процедуры обработки. При включении в конкретное контекстное окружение одна или несколько валентностей закрываются, в результате формируется сейрон – специализированное состояние семрона в конкретной ситуации.

Целью настоящей работы было описать базу семронов татарского языка как цифровой лингвистический ресурс, обеспечивающий верифицируемое представление лексико-грамматических единиц и их использование в объяснительном искусственном интеллекте. Фреймовое представление рассмотрено как способ организации записи семрона внутри БСБС. В работе модель цифрового ресурса, его структура, формат представления, а также представлены принципы верификации.

ОСНОВНАЯ ТЕРМИНОЛОГИЯ

Нами используется ограниченный набор терминов Семиотической генеративной распознающей модели (СГРМ) (см. табл. 1), которые необходимы для описания цифровой базы семронов.

Табл. 1. Основная терминология.

Термин	Рабочее определение
СГРМ	Семиотическая генеративная распознающая модель
БСБС	Большая семиотическая база семронов
Семрон	Потенциальная лексико-грамматическая смыслообразующая единица с формой, синтаксисом, семантикой, прагматикой и открытыми валентностями.
Сейрон	Специализированное состояние семрона в конкретной ситуации, где закрыта по крайней мере одна валентность.
Фрейм	Структурированная рабочая запись, включающая паспорт, слоты, фасеты, ограничения, валентности и процедуры.
Валентность	Потенциальная или реализованная связь семрона с другими семронами и элементами контекстного окружения.
Верифицируемая цепочка	Путь от формы и морфологического разбора к семрону, правилу, сейрону, слоту ситуационной модели и объяснению.

ЭЛЕКТРОННЫЕ ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ И ЗАДАЧА СЕМАНТИЧЕСКОГО СТРУКТУРИРОВАНИЯ ТАТАРСКОГО ЯЗЫКА

Электронные лингвистические ресурсы являются необходимым условием развития современных технологий обработки естественного языка. Для языков с развитой цифровой ресурсной базой такие ресурсы обычно представлены корпусами, словарями, морфологическими, синтаксическими и семантическими анализаторами, онтологиями и графами знаний. Для малоресурсных языков, к которым в цифровом отношении относится и татарский язык, такая инфраструктура развита существенно слабее. В работе [1] отмечено, что доля татарских текстов в Common Crawl составляет менее 0.01% общего объема данных.

Однако объем текстов сам по себе не решает задачу объяснимой обработки. Даже при наличии корпуса и морфологического анализатора система может определить состав словоформы, но не объяснить, какая смысловая функция реализуется конкретной морфемой в данной ситуации. Для татарского языка, где смысл часто формируется в аффиксальной цепочке, отмеченная проблема является особо острой.

Электронная библиотека в современном понимании не ограничивается хранением текстов. Постоянно подчеркивается, что электронные библиотеки включают не только текстовые и мультимедийные цифровые объекты, но и структурированные данные и описывающие их метаданные (например, [2]). В этом смысле цифровая база семронов может рассматриваться как специализированный электронный лингвистический ресурс, ориентированный на структурированное представление лексико-грамматического знания.

БСБС выполняет интегративную функцию (рис. 1). Она связывает форму, алломорфы, класс, синтаксические условия, семантические функции, прагматические условия, валентности, корпусные примеры, графовые связи и статус проверки в одной фреймовой записи. Поэтому такая база может использоваться не только как хранилище сведений, но и как рабочий слой объяснимой обработки текста.

источники лингвистических данных



Рис. 1. Место Большой семиотической базы семронов среди цифровых лингвистических ресурсов.

С точки зрения электронных библиотек такая постановка особенно важна, поскольку она включает метаданные, онтологии, концептуальное моделирование, графовые модели данных, извлечение знаний, семантическое структурирование контента и цифровое курирование данных [3]. БСБС является одним из ресурсов, предназначенным для семантического структурирования лексико-грамматического знания татарского языка.

СЕМИОТИЧЕСКАЯ УНИВЕРСАЛИЯ И СЕМРОН КАК ОСНОВА ФРЕЙМОВОГО ОПИСАНИЯ

В настоящей работе понятие семрона опирается на ранее разработанную концепцию семиотической универсалии. В этой концепции минимальной единицей описания лексико-грамматической структуры языка выступает не только слово или морфема как формальный элемент, а функционально-семиотическая единица, обладающая собственной структурой смыслообразования. Семиотическая универсалия может быть описана через имя, класс, синтаксис, семантику и прагматику, а ее ситуационная проактивность выражена в наличии потенциальных валентностей, активируемых в конкретном контекстном окружении [4].

Для целей цифровой реализации эта теоретическая единица получает операциональную форму семрона. Если семиотическая универсалия задает общий

теоретический уровень описания лексико-грамматической смыслообразующей единицы, то семрон является ее представлением в цифровой базе и процедурах обработки СГРМ.

Такое определение принципиально отличает семрон от обычной словарной единицы. Словарная статья фиксирует значение и употребление, но, как правило, не задает формальную структуру условий, при которых единица реализует ту или иную функцию. Семрон же должен содержать не только форму и толкование, но и набор возможных контекстных связей: синтаксических, семантических и прагматических.

Для агглютинативного татарского языка это представляет особый интерес. Значимая часть смысловой структуры словоформы формируется за счет аффиксальных цепочек, где отдельный показатель может реализовать разные функции в зависимости от основы, соседних аффиксов, дискурсивной памяти и ситуации. В [5] это показано на примерах аффиксальных морфем *-сы* и *-лар*.

Переход от семиотической универсалии к семрону можно представить как переход от теоретического описания к операциональной записи:

СУ → Semron → SemronRecord → Frame.

Первый уровень задает теоретическую единицу смыслообразования, второй – семрон как потенциальную лексико-грамматическую единицу. Третий уровень представляет семрон как запись в БСБС. Четвертый оформляет эту запись как фрейм, пригодный для обработки, проверки, специализации и связывания с другими элементами цифрового ресурса.

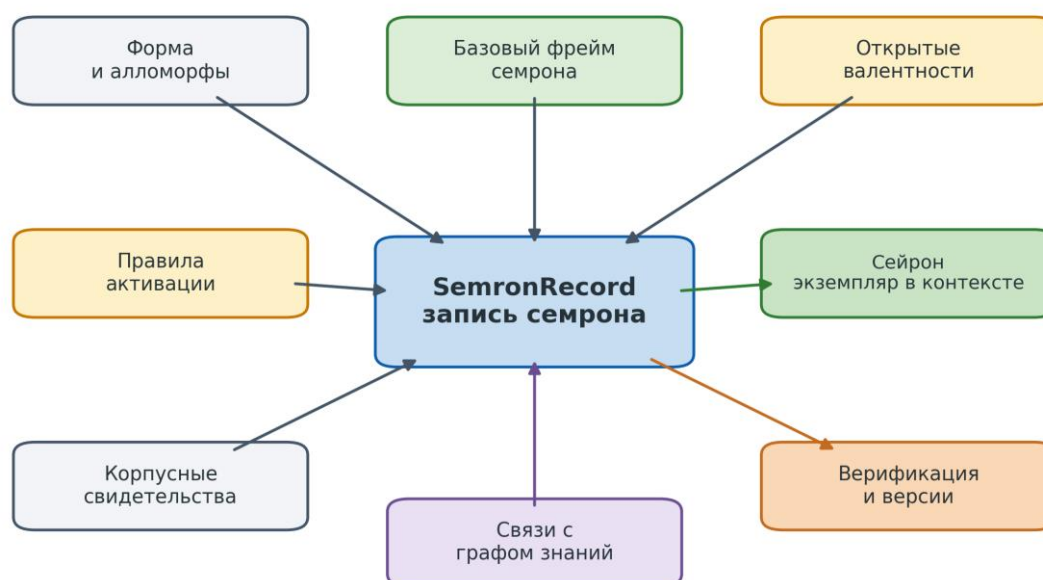
БСБС КАК ЦИФРОВОЙ ЛИНГВИСТИЧЕСКИЙ РЕСУРС

Большая семиотическая база семронов здесь рассматривается как цифровой лингвистический ресурс, предназначенный для представления лексико-грамматических единиц татарского языка в форме, пригодной для экспертного редактирования, машинной обработки и использования в СГРМ. В отличие от традиционного словаря, БСБС фиксирует не только значение единицы, но и ее потенциальные валентности, условия активации, корпусные примеры, связи с графом знаний и статус проверки.

Фреймовая организация БСБС позволяет объединить в одной записи разнотипные сведения. Паспорт записи содержит идентификатор, имя, версию и статус. Формальный блок фиксирует каноническую форму, алломорфы, класс и допустимые контексты. Содержательный блок описывает семантические функции и прагматические условия. Валентностный блок задает открытые валентности и условия их закрытия. Верификационный блок хранит корпусные примеры, связи с графом знаний, экспертные комментарии и статус проверки.

Такое устройство позволяет отличать потенциальную единицу от ее конкретной реализации. Базовый фрейм семрона описывает возможности единицы до включения в контекст. Когда семрон включается в конкретное контекстное окружение, часть его открытых валентностей может быть закрыта. В результате возникает сейрон – специализированное состояние семрона в конкретной ситуации.

БСБС должна быть курируемым ресурсом. Автоматические процедуры могут предлагать новые корпусные примеры, связи с графом знаний или варианты правил активации, но изменение статуса записи должно утверждаться экспертом. Это отличает БСБС от неуправляемых самообновляемых баз и делает ее пригодной для научной верификации.



Каждая запись связывает описание единицы, правила активации, внешние основания и состояние проверки.

Рис. 2. Логическая структура цифровой базы семронов.

На рис. 2 представлена логическая структура БСБС как база данных. Центральным объектом является запись семрона. С ней связаны форма и алломорфы, фреймовая структура, валентности, правила активации, корпусные примеры, связи с графом знаний, статусы проверки и экземпляры сейронов. Такая схема позволяет перенести обсуждение БСБС из уровня общей идеи в уровень проектируемой структуры базы.

ФРЕЙМОВАЯ СТРУКТУРА ЗАПИСИ СЕМРОНА

Фреймовая запись семрона является центральной единицей БСБС как цифрового лингвистического ресурса. Она должна одновременно выполнять две функции: фиксировать лингвистическое описание единицы и обеспечивать ее машинно-читаемое представление, поэтому запись семрона не может ограничиваться только формой, переводом или грамматическим тегом.

С учетом цели настоящей работы различать четыре уровня представления: сигнатуру семрона, запись семрона в БСБС, базовый фрейм семрона и специализированный фрейм сейрона (рис. 3). Сигнатура описывает лингвистическую единицу как потенциальный элемент языка; запись в БСБС расширяет ее служебными, верификационными и ресурсными полями; базовый фрейм фиксирует потенциальные возможности; специализированный фрейм описывает реализованное состояние в конкретной ситуации.

Сигнатуру семрона можно представить следующим образом:

$$s = \langle \text{name, class, form, syntax, semantics, pragmatics, V_open} \rangle.$$

Запись семрона в БСБС имеет вид

$$r(s) = \langle \text{ID, s, R, E_c, E_g, Q, Version} \rangle.$$

Базовый фрейм семрона может быть представлен так

$$F_base(s) = \{ \text{Passport, Slots, Facets, V_open, Rules, Procedures, Status} \}.$$

Специализированный фрейм сейрона имеет вид:

$$F_spec(seiron) = \langle F_base(s), C, V_closed, f, M, Chain \rangle.$$



Семрон хранит потенциальные связи; сейрон фиксирует реализованный смысловой узел.

Рис. 3. Базовый фрейм семрона и специализированный фрейм сейрона.

Фреймовая запись задает механизм перехода от потенциальной единицы к реализованной функции. Семрон хранит открытые валентности; контекстное окружение активирует одну или несколько из них; закрытая валентность переводит семрон в состояние сейрона; специализированный фрейм сейрона фиксирует реализованную функцию и ее место в ситуационной модели.

СЛОТЫ, ФАСЕТЫ И КОНТРОЛЬ КАЧЕСТВА ЗАПИСИ

Фреймовая структура записи семрона предполагает разделение между слотами и фасетами. Слотом называется поле фрейма или ситуационной модели, содержащее отдельный компонент смысла или формального описания. Фасет уточняет состояние слота: источник значения, ограничение, степень уверенности, экспертный статус, корпусное подтверждение, допустимость или дату проверки.

Использование фасетов позволяет превратить запись семрона из простой таблицы в контролируемый цифровой объект. Например, слот `semantic_functions` может содержать список возможных смысловых функций, а его фасеты фиксируют источник каждой функции, экспертную уверенность, наличие корпусного подтверждения и текущий статус проверки. Слот `open_valencies` может иметь фасеты `activation_condition`, `restriction`, `validation_status`; слот `corpus_evidence` может иметь фасеты `source_text`, `frequency`, `context_type`.

Для БСБС это принципиально важно, поскольку база должна быть не самоизменяющимся набором гипотез, а курируемым ресурсом. Каждая запись должна иметь статус, а каждое изменение – быть связано с версией, источником

и экспертным решением. Автоматические процедуры могут предлагать обновления, но не должны самостоятельно утверждать новую норму записи.

ВАЛЕНТНОСТЬ И ПЕРЕХОД ОТ СЕМРОНА К СЕЙРОНУ

Ключевым элементом фреймовой записи семрона является валентность, которая понимается как потенциальная или реализованная связь семрона с другими семронами и элементами контекстного окружения. Такая связь может быть синтаксической, семантической, прагматической или смешанной. Открытая валентность фиксирует возможность связи до включения семрона в конкретную ситуацию; закрытая – реализованную связь в конкретном контекстном окружении.



Проверяемые основания переходов: правило, корпусное свидетельство, связь с графом знаний или экспертный статус.

Рис. 4. Переход от записи семрона к объяснимой обработке текста.

Семрон хранит множество открытых валентностей. Например, аффиксальный семрон может иметь валентности связи с левой основой, правым аффиксом, дискурсивной памятью, корпусным употреблением или узлом графа знаний. До анализа конкретного текста эти валентности остаются потенциальными. Когда появляется контекстное окружение, одна или несколько валентностей могут быть закрыты (см. рис. 4).

Сейрон является результатом такой специализации. Он фиксирует не всю потенциальную структуру семрона, а конкретное реализованное состояние: какая валентность закрыта, какая функция активирована, какой слот ситуационной модели заполнен и какой путь проверки может быть предъявлен.

Рассмотрим примеры. Для семрона *-сЫ* могут быть предусмотрены функции принадлежности, намерения и субстантивации. В форме *бакчасы* закрыва-

ется валентность между владельцем и объектом принадлежности; в форме *бa-расы* активируется связь между субъектом и предполагаемым действием; в форме *атасы* возможна специализация через родственно-ролевой или прагматический контекст.

ОТ ФРЕЙМОВОЙ БАЗЫ К СЕМРОННОЙ ГИПЕРСЕТИ И ФОРМАТАМ ДАННЫХ

Фреймовая база семронов может рассматриваться как основа семронной гиперсети. При этом гиперсеть не является самостоятельным объектом хранения, независимым от БСБС. Она является результатом связывания базовых фреймов семронов и специализированных фреймов сейронов через валентности, правила активации, корпусные употребеления и связи с графом знаний.

Формально такую гиперсеть можно описать как

$$H = \langle S, F_b, F_s, V_o, V_c, R, E, Q \rangle,$$

где S – множество семронов, F_b – множество базовых фреймов семронов, F_s – множество специализированных фреймов сейронов, V_o – множество открытых валентностей, V_c – множество закрытых валентностей, R – правила активации, E – внешние основания, Q – статусы качества и проверки.

Для использования БСБС как электронного лингвистического ресурса необходимо поддерживать несколько форматов представления данных, соответствующих разным задачам работы с записью. Табличное представление предназначено для экспертного редактирования лингвистом; JSON или YAML – для программной обработки; RDF, OWL или GraphML – для связи с графом знаний и интеграции с онтологиями и внешними цифровыми библиотечными ресурсами. При этом речь идет не о дублировании ресурса, а о разных представлениях одной и той же записи.

ПРИМЕР ЗАПИСИ СЕМРОНА

В качестве основного примера рассмотрен семрон *-сЫ* (табл. 2), поскольку он не является центральным объектом первой статьи и уже описан в материалах по фреймовой модели семиотической универсалии. Для *-сЫ* зафиксированы алломорфы *-сы*, *-се*, класс «Аффиксальная морфема», а также функции принадлежности, намерения и субстантивации [5].

Табл. 2. Компактная запись семрона – *сbl*.

Поле	Значение
id	SEM_AFFIX_SY
name	- <i>сbl</i>
allomorphs	- <i>сы</i> , - <i>се</i>
class	аффиксальная морфема
semantic_functions	принадлежность; намерение; субстантивация
open_valencies	владелец ↔ объект; субъект ↔ предполагаемое действие; родственный термин ↔ ситуация обращения
examples	бакчасы; барасы; аласы; атасы
validation_status	экспертно описано / требует корпусной проверки

Машиночитаемое представление записи может быть дано в формате JSON. Ниже приведен сокращенный фрагмент, демонстрирующий принцип вложенной структуры.

```
{
  "id": "SEM_AFFIX_SY",
  "name": "-сbl",
  "allomorphs": ["-сы", "-се"],
  "class": "affixal_morpheme",
  "semantic_functions": ["POSS", "INTENT", "SUBST"],
  "open_valencies": [
    {"id": "V_SY_01", "type": "semantic", "description": "владелец – объект"},
    {"id": "V_SY_02", "type": "semantic_pragmatic", "description": "субъект – действие"},
    {"id": "V_SY_03", "type": "pragmatic", "description": "родственный термин – обращение"}
  ],
  "validation": {"status": "expert_described", "version": "0.1"}
}
```

Семрон -ЛАр использован только как компактная иллюстрация структуры записи в БСБС. Экспериментальная проверка его контекстных функций относится к отдельному этапу проекта и не является предметом настоящей работы.

```

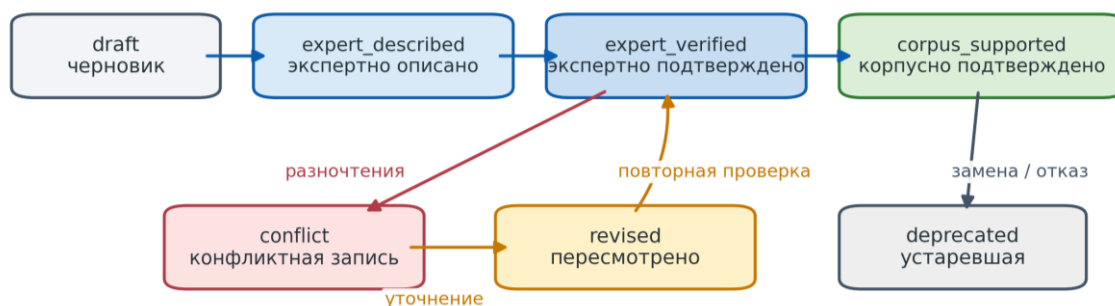
{
  "id": "SEM_AFFIX_LAR",
  "name": "-ЛАр",
  "allomorphs": ["-лар", "-ләр", "-нар", "-нәр"],
  "class": "affixal_polyfunctional_morpheme",
  "semantic_functions": ["quantitative_plurality", "associative_domain",
"collective_possessor"],
  "activation_rules": ["V_LAR_01", "V_LAR_02", "V_LAR_03"],
  "comment": "пример структуры записи; не экспериментальный результат
статьи"
}

```

ВЕРИФИКАЦИЯ И ЖИЗНЕННЫЙ ЦИКЛ БСБС

Верификация является обязательным элементом БСБС как цифрового ресурса. Поскольку запись семрона включает не только форму, но и семантические функции, прагматические условия, открытые валентности и правила активации, она должна проходить экспертную проверку и, по возможности, корпусное подтверждение.

Жизненный цикл записи (рис. 5) может включать несколько этапов. На первом этапе создается черновая запись, содержащая минимальный набор полей. На втором эксперт уточняет алломорфы, контексты, валентности и правила активации. На третьем запись получает статус экспертно описанной или экспертно подтвержденной. На четвертом к записи добавляются корпусные примеры и графовые связи. На пятом этапе запись может быть пересмотрена, если появятся новые данные или конфликтные интерпретации.



Автоматические процедуры рекомендуют изменения; статус записи утверждает эксперт.

Рис. 5. Жизненный цикл записи в Большой семиотической базе семронов.

Верификация БСБС должна быть многоуровневой. Экспертная верификация проверяет корректность лингвистического описания. Корпусная проверка подтверждает наличие употреблений. Граф знаний поддерживает типизацию сущностей и отношений. Верифицируемая цепочка показывает путь от формы к реализованной функции. В совокупности эти уровни делают БСБС пригодной для научного использования и последующей интеграции в СГРМ.

БОЛЬШАЯ СЕМИОТИЧЕСКАЯ БАЗА СЕМРОНОВ В АРХИТЕКТУРЕ СГРМ

В архитектуре СГРМ Большая семиотическая база семронов выполняет функцию ресурсного слоя. Она не заменяет все остальные компоненты системы, а предоставляет активируемые записи семронов, необходимые для перехода от языковой формы к смысловой структуре. При обработке текста входное выражение сначала проходит морфемную сегментацию и нормализацию. Затем соответствующие элементы сопоставляются с записями БСБС. Далее проверяются открытые валентности, условия их закрытия, контекстное окружение, корпусные примеры и связи с графом знаний.

Общий маршрут обработки можно представить так: входной текст → морфемная сегментация → БСБС → семроны → закрытие валентностей → сейроны → ситуационная модель → верифицируемая цепочка → объяснение. Этот маршрут не должен пониматься как описание готовой полнофункциональной системы. Здесь он использован для показа места БСБС в общей архитектуре СГРМ.

ОБСУЖДЕНИЕ

Предложенная модель БСБС решает несколько задач. Во-первых, она задает структуру цифрового ресурса, способного представлять лексико-грамматические единицы татарского языка в машинно-читаемом виде. Во-вторых, она позволяет связать форму, семантику, прагматику, валентности и корпусные употребления в одной записи. В-третьих, вводит механизм перехода от потенциальной единицы к реализованному смысловому узлу через закрытие валентностей.

Основное отличие БСБС от традиционных ресурсов состоит в том, что она не ограничивается хранением данных, а задает условия смысловой активации. Словарь сообщает значение, корпус показывает реальное употребление в языке,

морфологический анализатор выделяет форму, а граф знаний фиксирует предметные связи. БСБС связывает эти уровни и обеспечивает возможность построения верифицируемой цепочки.

Предложенная модель имеет и ограничения. Создание БСБС требует значительной экспертной работы; для каждой записи необходимо определить статусы проверки и версионирование; корпусные употребления для отдельных семронов могут быть недостаточны, особенно если речь идет о редких прагматических функциях; требуется разработка инструментов для согласованного редактирования табличной, машиночитаемой и графовой форм записи.

Перспективы работы связаны с расширением набора семронов, подготовкой стандарта записи, формированием набора экспертно проверенных примеров и созданием интерфейса для редактирования БСБС. Отдельной задачей является интеграция БСБС с существующими цифровыми ресурсами татарского языка, включая корпуса, словари, онтологические платформы и графы знаний.

ЗАКЛЮЧЕНИЕ

Предложена модель базы семронов татарского языка как цифрового лингвистического ресурса. Показано, что БСБС не сводится к словарю, корпусу, морфологическому анализатору, графу знаний или базе правил, а выполняет интегративную функцию: связывает форму, алломорфы, класс, синтаксис, семантику, прагматику, валентности, корпусные примеры, графовые связи и статусы проверки в одной фреймовой записи.

Основной единицей БСБС является семрон – потенциальная лексико-грамматическая смыслообразующая единица с открытыми валентностями. Для его описания предложена фреймовая структура, включающая паспорт, слоты, фасы, открытые валентности, правила активации, процедуры обработки и статус проверки. При включении семрона в конкретное контекстное окружение одна или несколько валентностей закрываются и формируется сейрон – специализированное состояние семрона в конкретной ситуации.

Предложенное различие базового фрейма семрона и специализированного фрейма сейрона позволяет описать переход от потенциальной единицы к реализованному смысловому узлу. Множество таких записей образует основу

семронной гиперсети и может использоваться в СГРМ для объяснимой обработки татарского текста.

Практическая значимость работы состоит в формировании модели базы семронов как цифрового ресурса, пригодного для экспертного наполнения, машинной обработки, корпусной проверки, связи с графом знаний, версионирования и последующего использования в объяснительном искусственном интеллекте. Дальнейшая работа будет направлена на расширение набора семронов, уточнение форматов представления, разработку процедур верификации и подготовку пилотной версии БСБС как курируемого электронного лингвистического ресурса татарского языка.

Благодарности

Работа выполнена в рамках Соглашения №13/2025-ПД от 22.12.2025 о предоставлении гранта научным и научно-педагогическим работникам обособленных структурных подразделений Академии наук Республики Татарстан с целью стимулирования их к защите докторских диссертаций и выполнению научно-исследовательских работ.

СПИСОК ЛИТЕРАТУРЫ

1. *Alimova I., Monogov B., Mazur A., Antonov D., Karimov V., Egorov V., Khamimov B., Panchenko A.* The Case of Tatar // Proceedings of the 15th Joint Conference on Lexical and Computational Semantics. San Diego, California: Association for Computational Linguistics, 2026. P. 264–274.

URL: <https://aclanthology.org/2026.starsem-conference.17> (дата обращения: 02.06.2026). <https://doi.org/10.48550/arXiv.2606.26015>

2. Электронные библиотеки. О журнале. URL: <https://rdl-journal.ru/about> (дата обращения: 01.06.2026).

3. Электронные библиотеки. Тематика журнала.
URL: <https://rdl-journal.ru/about/themes> (дата обращения: 01.06.2026).

4. *D. Suleymanov, A. Gilmullin and I. Mukhametzyanov.* Semiotic Universal as an Atomic Element of the Lexical-Grammatical Structure of the Tatar Language, 2025 10th International Conference on Computer Science and Engineering (UBMK), Istanbul, Turkiye. 2025. P. 1767–1771.

<https://doi.org/10.1109/UBMK67458.2025.11206836>.

5. *Suleymanov D.Sh., Gilmullin R.A., Fridman A.Ya., Mukhamedshin D.* Semantic Models of Agglutinative Languages // 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE). 2024. P. 1930–1935. <https://doi.org/10.1109/PIERE62470.2024.10804991>

6. *Сулейманов Д.Ш., Мухамедшин Д.Р., Гильмуллин Р.А., Фридман А.Я.* Семиотическая модель для представления когнитивных структур языка агглютинативного типа // Искусственный интеллект и принятие решений. 2024. № 4. С. 55–66. <https://doi.org/10.14357/20718594240405>

7. *Stefanuk V.L., Taisina E.A., Suleymanov Dz.Sh., Fridman A.Ya.* (Eds.). Semiotic Aspects of Modeling the Cognitive System for the Artificial Intelligence. Editions du JIPTO, 2024. 330 p.

8. *Тарасов В.Б.* Семиотические аспекты коммуникации между агентами в мультиагентных системах // Semiotic Aspects of Modeling the Cognitive System for the Artificial Intelligence. Editions du JIPTO, 2024. С. 155–188.

9. *Туган тел: национальный корпус татарского языка.*
URL: <https://tugatel.tatar/> (дата обращения: 02.06.2026).

10. *Minsky M.* A Framework for Representing Knowledge // The Psychology of Computer Vision / ed. by P.H. Winston. New York: McGraw-Hill, 1975. P. 211–277.

11. RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation. URL: <https://www.w3.org/TR/rdf11-concepts/> (дата обращения: 30.05.2026).

12. OWL 2 Web Ontology Language Document Overview. W3C Recommendation. URL: <https://www.w3.org/TR/owl2-overview/> (дата обращения: 20.05.2026).

THE TATAR LANGUAGE SEMRON DATABASE AS A DIGITAL LINGUISTIC RESOURCE FOR EXPLAINABLE ARTIFICIAL INTELLIGENCE

R. A. Gilmullin^[0000-0002-8520-8921]

Tatarstan Academy of Sciences, Kazan, Russia

rinatgilmullin@gmail.com

Abstract

The article considers the Tatar language semron database as a digital linguistic resource for explainable artificial intelligence. A model of the Big Semiotic Database of Semrons is proposed, in which lexical-grammatical units are represented as semrons, i.e. potential meaning-forming units with form, class, syntactic, semantic and pragmatic characteristics, valencies, activation rules, corpus evidence, knowledge graph links and expert verification statuses. Frame representation is used as a way of organizing a semron record: the base frame fixes the potential capabilities of a unit, while the specialized seiron frame describes its contextual realization after one or more valencies have been closed. It is shown that such a database can function as a resource layer of the Semiotic Generative Recognizing Model, supporting the transition from the morphemic structure of Tatar text to a situational model, a verifiable chain and an explanation. Compact record examples demonstrate the frame structure, types of slots and facets, verification principles and possible data representation formats.

Keywords: *semron, semron database, Tatar language, digital linguistic resource, frame model, valency, seiron, knowledge graph, corpus layer, expert verification, explainable artificial intelligence, Semiotic Generative Recognizing Model.*

REFERENCES

1. Alimova I., Monogov B., Mazur A., Antonov D., Karimov V., Egorov V., Khakimov B., Panchenko A. The Case of Tatar // Proceedings of the 15th Joint Conference on Lexical and Computational Semantics. San Diego, California: Association for Computational Linguistics, 2026. P. 264–274. URL: <https://aclanthology.org/2026.starsem-conference.17> (accessed 02.06.2026).
<https://doi.org/10.48550/arXiv.2606.26015>
2. Russian Digital Libraries Journal. About the journal. Available at:

<https://rdl-journal.ru/about> (accessed 06.06.2026).

3. Russian Digital Libraries Journal. Topics. Available at:
<https://rdl-journal.ru/about/themes> (accessed 06.06.2026).

4. *D. Suleymanov, A. Gilmullin and I. Mukhametzyanov. Semiotic Universal as an Atomic Element of the Lexical-Grammatical Structure of the Tatar Language*, 2025. 10th International Conference on Computer Science and Engineering (UBMK), Istanbul, Turkiye, 2025. P. 1767–1771.

<https://doi.org/10.1109/UBMK67458.2025.11206836>

5. *Suleymanov D.Sh., Gilmullin R.A., Fridman A.Ya., Mukhamedshin D. Semantic Models of Agglutinative Languages // 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE). 2024. P. 1930–1935.*
<https://doi.org/10.1109/PIERE62470.2024.10804991>

6. *Suleymanov D.Sh., Mukhamedshin D.R., Gilmullin R.A., Fridman A.Ya. Semiotic model for representing cognitive structures of an agglutinative language. Iskusstvennyj intellekt i prinyatie reshenij. 2024. No. 4. P. 55–66 (In Russ.).*

7. *Stefanuk V.L., Taisina E.A., Suleymanov Dz.Sh., Fridman A.Ya., eds. Semiotic Aspects of Modeling the Cognitive System for the Artificial Intelligence. Editions du JIPTO; 2024. 330 p.*

8. *Tarasov V.B. Semiotic aspects of communication between agents in multi-agent systems. In: Semiotic Aspects of Modeling the Cognitive System for the Artificial Intelligence. Editions du JIPTO. 2024. P. 155–188 (In Russ.).*

9. *Tugan Tel: National Corpus of the Tatar Language. Available at: <https://tugantel.tatar/> (accessed 01.06.2026). (In Russ.)*

10. *Minsky M. A Framework for Representing Knowledge. In: Winston P.H., ed. The Psychology of Computer Vision. New York: McGraw-Hill. 1975. P. 211–277.*

11. *RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation. Available at: <https://www.w3.org/TR/rdf11-concepts/> (accessed 30.05.2026).*

12. *OWL 2 Web Ontology Language Document Overview. W3C Recommendation. Available at: <https://www.w3.org/TR/owl2-overview/> (accessed 23.05.2026).*

СВЕДЕНИЯ ОБ АВТОРЕ



ГИЛЬМУЛЛИН Ринат Абрекович – кандидат физико-математических наук, директор Института прикладной семиотики Академии наук Республики Татарстан. Область научных интересов: искусственный интеллект, прикладная семиотика, обработка естественного языка, цифровые лингвистические ресурсы.

Rinat Abrekovich GILMULLIN – Candidate of Physical and Mathematical Sciences, Director of the Institute of Applied Semiotics of the Tatarstan Academy of Sciences. Research interests: artificial intelligence, applied semiotics, natural language processing, digital linguistic resources.

e-mail: rinatgilmullin@gmail.com

ORCID: 0000-0002-8520-8921

Материал поступил в редакцию 30 июня 2026 года