

АРХИТЕКТУРА ЛИНГВИСТИЧЕСКОЙ ПЛАТФОРМЫ “TurkLang”

А. Р. Гатиатуллин^[0000-0003-3063-8147]

Академия наук Республики Татарстан, г. Казань, Россия

ayrat.gatiatullin@gmail.com

Аннотация

Описана архитектура многофункциональной лингвистической платформы “TurkLang”, создаваемой в Институте прикладной семиотики Академии наук Республики Татарстан. В основе системы лежит трехуровневая база знаний: языковые универсалии, описания единиц конкретных языков и эмпирические речевые данные. База знаний реализована в виде графа, где основной единицей хранения выступает морфема. Это решение обусловлено элементарно-комбинаторной природой тюркских языков: словоформа представляет собой линейную цепочку, состоящую из корня и последовательно присоединенных аффиксов – модификаторов, каждый из которых выражает одну грамматическую категорию (падеж, число, время и т. д.) в строго фиксированном порядке. Рассмотрены уровни абстракции, типы связей между сущностями, механизмы интеграции с внешними источниками (электронная энциклопедия, лексикографические ресурсы, геоинформационные системы). Особое внимание уделено применению международных стандартов синтаксической, семантической и прагматической (DiAML) аннотаций. Выполнено сравнение с существующими платформами (Sketch Engine, LingvoDoc, Apertium). На сегодняшний день база знаний платформы охватывает более 30 языков и диалектов, содержит более 237 тыс. корневых морфем и около 40 тыс. правил сочетаемости, а корпусная часть насчитывает свыше 180 млн словоупотреблений.

Ключевые слова: лингвистическая платформа, граф знаний, тюркские языки, морфемоцентричность, трехуровневая архитектура, Универсальные зависимости (Universal Dependencies), Универсальное смысловое представление (Universal Meaning Representation), агглютинативный язык, семиотическая вертикаль.

ВВЕДЕНИЕ

Обработка естественного языка для тюркских языков сопряжена с двумя фундаментальными проблемами.

Первая проблема является структурной. Тюркские языки по своей структуре можно отнести к элементно-комбинаторному типу [1] (в лингвистике их называют агглютинативными).

Для тюркских языков справедливо следующее: словоформа всегда строится по единой схеме, где каждый аффикс последовательно добавляется к основе и несет одну грамматическую характеристику. С формальной точки зрения такая морфологическая структура может быть описана регулярной грамматикой и моделируется конечным автоматом [2]. Однако большинство существующих инструментов NLP, включая синтаксические парсеры и стандарты аннотации (Universal Dependencies), разрабатывалось для индоевропейских языков и не учитывают структурно-функциональные особенности тюркских языков [3]. В результате некоторый объем информации в таких моделях теряется, что снижает качество автоматического анализа.

Вторая проблема – это ресурсная. Подавляющее большинство тюркских языков (за исключением турецкого) являются малоресурсными [4]. Проявляется это в нехватке полноценных электронных словарей, электронных корпусов с разметкой на различных языковых уровнях, лингвистических процессоров и унифицированных лингвистических стандартов. Существующие цифровые ресурсы создаются разрозненными коллективами, используют несовместимые форматы и системы тегов, что делает их интеграцию практически невозможной.

Для преодоления этих барьеров в Институте прикладной семиотики Академии наук Республики Татарстан (г. Казань) разрабатывается платформа “TurkLang”. Ее архитектура строится на трехуровневой модели данных, где центральным элементом является морфема, а не слово, как во многих известных моделях. Это позволяет:

- явно фиксировать порядок аффиксов в словоформе;
- представлять структуру словоформы, предложения и их значения в виде графа знаний;

- получать значение лингвистической единицы или конструкции путем обхода графа.

Ниже описано архитектурное решение: организация данных, уровни абстракции, интеграционные механизмы с внешними источниками, технологический стек, а также результаты практической реализации и сравнение с существующими аналогами.

1. АНАЛИЗ СУЩЕСТВУЮЩИХ РЕШЕНИЙ

Для объективной оценки архитектурных решений платформы “TurkLang” необходимо рассмотреть ее в контексте существующих лингвистических платформ, графов знаний и цифровых ресурсов, ориентированных на тюркские и другие агглютинативные языки.

1.1. Корпусные платформы

Sketch Engine [5] – одна из наиболее распространенных корпусных платформ, разработанная компанией Lexical Computing. Система предоставляет доступ к более чем 800 корпусам на 100+ языках, включая шесть тюркских языков (турецкий, казахский, узбекский, киргизский, азербайджанский, татарский).

Ключевая особенность Sketch Engine заключается в наличии богатого набора инструментов анализа корпусов: конкорданс, частотные списки, коллокации, *n*-граммы, тематические списки ключевых слов, автоматического аннотирования, создания подкорпусов. API для разработчиков предоставляет программный интерфейс для интеграции с внешними системами.

Система Sketch Engine является менеджером корпусов, а не платформой для формального описания языка: в ней отсутствуют онтологии, грамматические правила и семантические модели. Кроме того, она не фиксирует порядок морфем в словоформе, что критично для агглютинативных языков.

1.2. Платформы для документирования языков

LingvoDoc [6] – это платформа для совместного документирования языков, разработанная в Институте системного программирования РАН совместно с Институтом языкознания РАН. Система ориентирована на полевых лингвистов и исследователей, работающих с малоресурсными и исчезающими языками.

Ключевой особенностью платформы является возможность коллективной работы с языковыми материалами, поддержкой распределенной обработки и сохранением прав создателей словарей и корпусов. В платформе реализованы многослойные словари. Инструменты анализа включают модули для фонетического, морфологического и этимологического анализов.

Как и в Sketch Engine, в этой платформе отсутствует семантический слой, что ограничивает возможности глубоких компаративных исследований по содержанию, а не только по форме. Платформа ориентирована на хранение и первичный анализ полевых данных, но не предоставляет формальных грамматических моделей для автоматической обработки.

1.3. Системы машинного перевода

Apertium [7] – это открытая платформа машинного перевода, основанная на правилах (rule-based). Изначально разрабатывалась для перевода между близкородственными языками.

Ключевой особенностью платформы является использование правил на основе конвейерной обработки. Платформа реализована на базе свободного программного обеспечения под лицензией GNU GPL, не требует API-ключей или аутентификации. Каждая языковая пара разрабатывается изолированно, без общей базы знаний для всех тюркских языков. Не реализована единая онтологическая модель для тюркских языков. Популярность платформы снизилась с развитием нейросетевых технологий машинного перевода, которые превзошли по качеству системы машинного перевода типа Apertium, основанных на правилах. Однако сервисы указанной платформы могут быть использованы для решения других задач, например, в качестве механизма аугментирования данных для малоресурсных тюркских языков.

1.4. Семантические ресурсы

KeNet [8] (турецкий WordNet) является наиболее полным семантическим ресурсом для турецкого языка, созданный методом «снизу вверх» на основе толкового словаря Турецкого лингвистического общества. Содержит 76757 синсетов, имеет как внутриязыковые семантические отношения, так и связи с Princeton WordNet через межъязыковые отношения.

Анализ интернет-ресурсов показал, что из тюркских языков WordNet-ресурсы также реализованы для узбекского и уйгурского языков: UZ-WORDNET [9] и Uyghur WordNet. UZ-WORDNET создан с помощью метода автоматического извлечения концептов из английского WordNet. Других специализированных WordNet-ресурсов для тюркских языков не обнаружено. Существующие ресурсы (кроме KeNet) созданы автоматическими методами и требуют ручной верификации.

Turkish FrameNet [10] – это фреймовый ресурс для турецкого языка, охватывающий 8 доменов с наиболее частотными предикатами: Activity, Cause, Change, Motion, Cognition, Perception, Judgement, Commerce. Содержит 2700 вручную аннотированных примеров и 19221 токенов.

Uyghur FrameNet [11] является фреймовым ресурсом для уйгурского языка.

Для остальных тюркских языков фреймовые ресурсы не найдены.

Framester – это концентратор, связывающий FrameNet, WordNet, VerbNet, BabelNet [12], DBpedia, Yago и DOLCE-Zero [13]. Поддерживает OWL-рассуждения на основе фреймовой семантики, но ориентирован преимущественно на английский язык.

В последние годы активно развиваются формальные языки для представления смысла целых предложений и текстов. Одним из наиболее перспективных стандартов разметки в этой области является Uniform Meaning Representation (UMR). UMR – это графовый формализм для представления смысла, который пришел на смену стандарту Abstract Meaning Representation (AMR). В отличие от AMR, изначально разработанного для английского языка, UMR с самого начала создавался как межкузыковой стандарт, применимый к типологически разнообразным языкам, включая агглютинативные.

Ключевой особенностью UMR является то, что смысл текста в нем представляется в виде ориентированного графа, узлами которого являются концепты (действия, объекты, атрибуты), а ребрами – семантические отношения.

Пример UMR-графа для татарского предложения «*Малай урманга барды*»:

```
(e / go-01
  :ARG1 (m / малай)
  :goal (y / урман :number plural)
  :time past
)
```

В отличие от AMR, который позволяет описывать только семантику отдельных предложений, UMR позволяет представлять семантические взаимосвязи между предложениями.

В UMR появились механизмы, которые позволяют описывать значения таких категорий, как эвиденциальность (источник информации), аспектуальность, категории лица и числа, что делает UMR особенно релевантным для тюркских языков. Для семиотической вертикали, представленной в настоящей работе, особую ценность представляет возможность формального отображения между синтаксическими структурами в формате Universal Dependencies и семантическими ролями в UMR.

1.6. Синтаксические аннотации: universal dependencies

В рамках международного проекта Universal Dependencies [14] (UD) созданы древовидные банки для 11 тюркских языков. В релизе UD v2.16 (2025 г.) представлены первые полностью выровненные параллельные UD-деревья для четырех тюркских языков: азербайджанского, киргизского, турецкого и узбекского.

Ключевая проблема стандарта UD для тюркских языков заключается в том, что в нем нет возможности фиксировать порядок аффиксов внутри словоформы. Для языков с элементарно-комбинаторной структурой, к числу которых относятся и тюркские языки, это важно, поскольку, например, турецкие слова:

masadakilerin = masa + LOC(da) + REL (ki) + PL(ler)+ GEN(in),

masaninkilerde = masa + GEN (nin) + REL(ki) + PL(ler)+ LOC(de)

имеют одинаковый набор грамматических признаков в стандартном UD, но разный порядок аффиксов, что задает различные смыслы и синтаксические структуры. Первое слово имеет значение «принадлежащий тем, которые находятся на столе», а второе «находящиеся у тех, которые принадлежат столу». Это приводит к тому, что появляются различные синтаксические и семантические структуры с одинаковым набором грамматических параметров.

2. БАЗОВЫЕ ПРИНЦИПЫ АРХИТЕКТУРЫ ПЛАТФОРМЫ “TurkLang”

В основе архитектуры платформы лежат четыре базовых взаимосвязанных принципа.

1. Трехуровневая организация определяет способ хранения данных.
2. Многоуровневая согласованность (семиотическая вертикаль) задает систему взаимосвязей между языковыми уровнями и определяет, какие данные и в каком формате должны быть описаны.
3. Морфемоцентричность выбирает базовую единицу хранения.
4. Параметризация обеспечивает масштабирование на все языки семьи через механизм наследования.

Вместе они определяют способ хранения, обработки и пополнения лингвистической информации.

2.1. Принцип трехуровневой организации базы знаний

База знаний платформы разделена на три логических уровня, каждый из которых отвечает за свой тип лингвистической информации (рис. 1). Связи направлены от конкретного к абстрактному (уровень 3 → уровень 2 → уровень 1), что исключает циклические зависимости и упрощает поддержку и масштабирование системы.

В основе этой организации лежит семиотическая вертикаль – модель языка как многоуровневой знаковой системы, где каждый уровень опирается на некоторый международный стандарт. Однако сама вертикаль является частью онтологического ядра (уровень 1), а не самостоятельной надстройкой над архитектурой.



Рис. 1. Трехуровневая организация базы знаний.

2.1.1. Уровень 1: Онтологическое ядро, лингвистические универсалии

Уровень 1 содержит языконезависимые сущности. Он образует онтологический каркас платформы – систему понятий и отношений, на которую ссылаются все остальные уровни. Состав уровня 1 приведен далее.

Семантическая сеть на основе WordNet – это иерархия концептов с отношениями «род-вид» (гиперонимия/гипонимия), «часть-целое» (меронимия) и другими. Исходно транслирован из Принстонского WordNet, дополнен и продолжает дополняться тюркскими культурными концептами, такими как национальные блюда, одежда, праздники, родственные отношения и др. Содержит около 60 тыс. концептов.

Библиотека ситуаций на основе FrameNet. В отличие от WordNet, который описывает статическую лексическую семантику, т. е. отношения между словами, FrameNet описывает динамические ситуации – фреймы (например, «Передача», «Движение», «Восприятие») с набором участников – семантических ролей.

Фреймом в компьютерной лингвистике является схематическое представление типовой ситуации (события, состояния, отношения). Например, фрейм «Передача» включает участников «Дающий», «Получатель» и «Передаваемый объект».

Семантические роли – участники ситуации. Они делятся на обязательные (Core) – без них фрейм не может быть полноценно представлен, и факультативные (Non-Core) – могут присутствовать или отсутствовать в зависимости от контекста.

Лексическими единицами выступают слова, которые активируют фрейм (например, глаголы *дать*, *подарить*, *вручить* активируют фрейм «Передача»).

В составе онто-лингвистической платформы FrameNet обеспечивает ситуационную семантику – описание того, как слова организуются вокруг типовых ситуаций. Это позволяет:

- связывать глагольные морфемы с фреймами, эта связка называется активацией фрейма (отношение *activatesFrame*);
- описывать валентностные структуры: какие участники и в каких ролях могут появляться при данном глаголе;
- сопоставлять способы выражения одной и той же ситуации в разных тюркских языках;
- разрешать лексическую многозначность: один глагол может активировать разные фреймы в различных контекстах.

Грамматические категории и граммемы – это системы противопоставленных грамматических значений.

Грамматическая категория представляет собой систему грамем, объединенных на основе общности семантического признака. Так, категория падежа включает граммемы Nom, Gen, Acc, Dat, Loc, Abl, каждая из которых является частным значением внутри этой системы. Наряду с ними выделяются квазиграммемы, фиксирующие оппозицию по типу «наличие/отсутствие признака». Примером может служить уподобительное значение, выражаемое аффиксами *-сыман* или *-дай*; такие признаки не складываются в полноценную грамматическую категорию.

В рамках платформы грамматические категории и граммемы выполняют роль связующего звена между морфологическим и семантическим уровнями.

Аффиксальные морфемы связываются с граммемами через отношение выражения (*expresses*); грамлемы, в свою очередь, транслируются в семантические роли в соответствии с параметрами отображения, задаваемыми компонентом D, который описан далее. Кроме того, они участвуют в формировании синтаксических правил на основе UD, определяя, какие словоформы допустимы в тех или иных синтаксических позициях.

Для единообразной записи произношения использован Международный фонетический алфавит (IPA), представляющий собой стандартизованную систему знаков, где каждый символ жестко соответствует определенному звуку или классу звуков независимо от конкретного языка. В нашей платформе IPA используются для фонологической разметки корпусных данных, что позволяет сопоставлять произносительные варианты одного слова на различных диалектах и языках, а также фиксировать аутентичное звучание для языков, находящихся под угрозой исчезновения.

Прагматический слой описан с помощью категорий DiAML, которые включают измерения, коммуникативные функции, квалификаторы и отношения, регламентированные стандартом ISO 24617-2.

2.1.2. Уровень 2: Грамматическое описание языков

Уровень 2 содержит зависимые от конкретного языка сущности: это инвентарь единиц и правила их комбинаторики для каждого языка или диалекта. Он является ядром платформы с точки зрения формализации конкретных языков.

Блок А: Базовые лингвистические единицы (инвентарь языка)

Фонемы – звуковой инвентарь языка с привязкой к IPA-символам уровня 1.

Алломорфы – конкретные фонетические варианты морфем с условиями выбора (сингармонизм, ассимиляция). У каждого алломорфа есть форма, тип (корневой, аффиксальный, аналитический), признак терминальности (может ли стоять последним в словоформе).

Морфемы – подразделяются на:

- корневые – несут лексическое значение, связаны с концептами 1-го уровня платформы;

- аффиксальные – выражают грамматические значения, связаны с граммами уровня 1;
- аналитические – отдельно стоящие служебные слова (послелог, частица, вспомогательные глаголы), функционально эквивалентные аффиксам, но пишущиеся раздельно;
- устойчивые выражения (MWE) – многословные конструкции с некомпозиционным значением, хранятся как отдельные лексические единицы со ссылкой на синсет и фрейм уровня 1.

Блок Б: Правила комбинаторики (грамматика языка)

Отношение Next – ориентированный граф допустимых последовательностей алломорфов в словоформе. Правила заданы на уровне конкретных алломорфов, что позволяет учитывать сингармонизм и ассимиляцию.

UD-правила описывают синтаксические зависимости: какие части речи и граммы могут быть подлежащим (nsubj), дополнением (obj), обстоятельством (obl) и т. д., с учетом порядка слов в предложении по правилу SOV.

Для решения проблемы фиксации порядка морфем в платформе разработано неинвазивное расширение стандарта UD – TurkUD. Оно сохраняет порядок морфем в поле MISC формата CONLL-U с использованием формата

$$\text{MorphLattice} = \text{урман}[\text{Stem}] | \text{нар}[\text{Plur}] | \text{га}[\text{Dat}].$$

Это расширение не нарушает совместимость со стандартным UD: парсеры, не поддерживающие TurkUD, игнорируют поле MISC, а поддерживающие получают дополнительную информацию о порядке морфем.

UMR-правила отображают синтаксические зависимости в семантические роли (ARG0–ARG5, :goal, :source, :time и т. д.) в зависимости от активированного фрейма и контекста.

В рамках платформы “TurkLang” формализм UMR использован на уровне эмпирических данных (слой L5 корпуса, описан далее). Аннотация строится на основе правил отображения, хранящихся в уровне 2 (компонент D параметрической модели). Процесс включает три этапа.

Синтаксический анализ. На основе UD-дерева предложения определяются синтаксические зависимости между словоформами (например, nsubj для подлежащего, obl:dat для косвенного дополнения в дательном падеже).

Отображение в семантические роли. Для каждой синтаксической зависимости применяются правила компонента D, которые сопоставляют грамлеммы и синтаксические роли с семантическими ролями UMR. Например:

nsubj → :ARG0 (агенса/экспериментатор)

obj → :ARG1 (пациент/тема)

obl:dat → :goal (в контексте фрейма движения),
 :recipient (в контексте фрейма передачи)

obl:loc → :location

obl:abl → :source

Выбор конкретной семантической роли для одной и той же грамлеммы (например, дательного падежа) зависит от активированного глаголом фрейма FrameNet. Так, в предложении «Малай урманга барды» (фрейм MOTION) аффикс -ГА отображается в роль :goal, а в предложении «Мин сиңа китап бирдем» (фрейм TRANSFER) – в роль :recipient.

Остановимся на построении UMR-графа. На основе определенных семантических ролей и атрибутов (время, аспект, эвиденциальность) формируется ориентированный граф, где узлы соответствуют концептам (событиям, сущностям), а ребра – семантическим отношениям (пример приведен выше).

Такой подход обеспечивает детерминированную и масштабируемую семантическую аннотацию корпуса, поскольку все правила отображения заданы явно и могут быть параметризованы для каждого языка (компонент D). Кроме того, UMR-аннотация используется на этапе верификации при генерации текста нейросетевыми моделями: агент-семантик проверяет, соответствует ли сгенерированный UMR-граф правилам платформы, и при обнаружении несоответствия инициирует регенерацию или формирует объяснение ошибки.

DiAML-правила сопоставляют грамматические формы (вопросительные частицы, эвиденциальные аффиксы, наклонения) с прагматическими функциями и квалификаторами.

2.1.3. Уровень 3: Эмпирические данные

Уровень 3 содержит реальные языковые данные (корпусы) с многослойной аннотацией, а также географическую привязку для диалектологических ис-

следований. Он распределен между графовой СУБД (для аннотаций), реляционной СУБД (для метаданных и геоданных) и объектным хранилищем (для мультимедиа).

Многослойное представление корпуса

Корпус проходит последовательные этапы обработки, каждый из которых добавляет новый слой аннотации. Переход между слоями L0–L6 автоматизирован через конвейер многоуровневой аннотации корпуса (рис. 2).

Каждый слой ссылается на правила из уровня 2 и универсалии из уровня 1. Токены хранятся как пути в графе Next, что позволяет выполнять быстрые обходы последовательностей морфем и семантический поиск по морфемной структуре.

Геолингвистические данные

Для поддержки диалектологических и ареальных исследований в уровень 3 интегрированы геолингвистические данные:

- Locality – населенный пункт с координатами по стандарту WGS84 и административной принадлежностью;
- Informant – информант с биографическими и социолингвистическими данными (год рождения, пол, образование, место рождения, миграционная история);
- Expedition – полевая экспедиция с метаинформацией (даты, состав исследователей, оборудование);
- Document – привязка документов к населенному пункту, информанту и экспедиции;
- Area – ареал распространения идиома (полигон на карте);
- Isogloss – изоглосса для лингвистического параметра, строится автоматически на основе значений параметров в точках сбора.

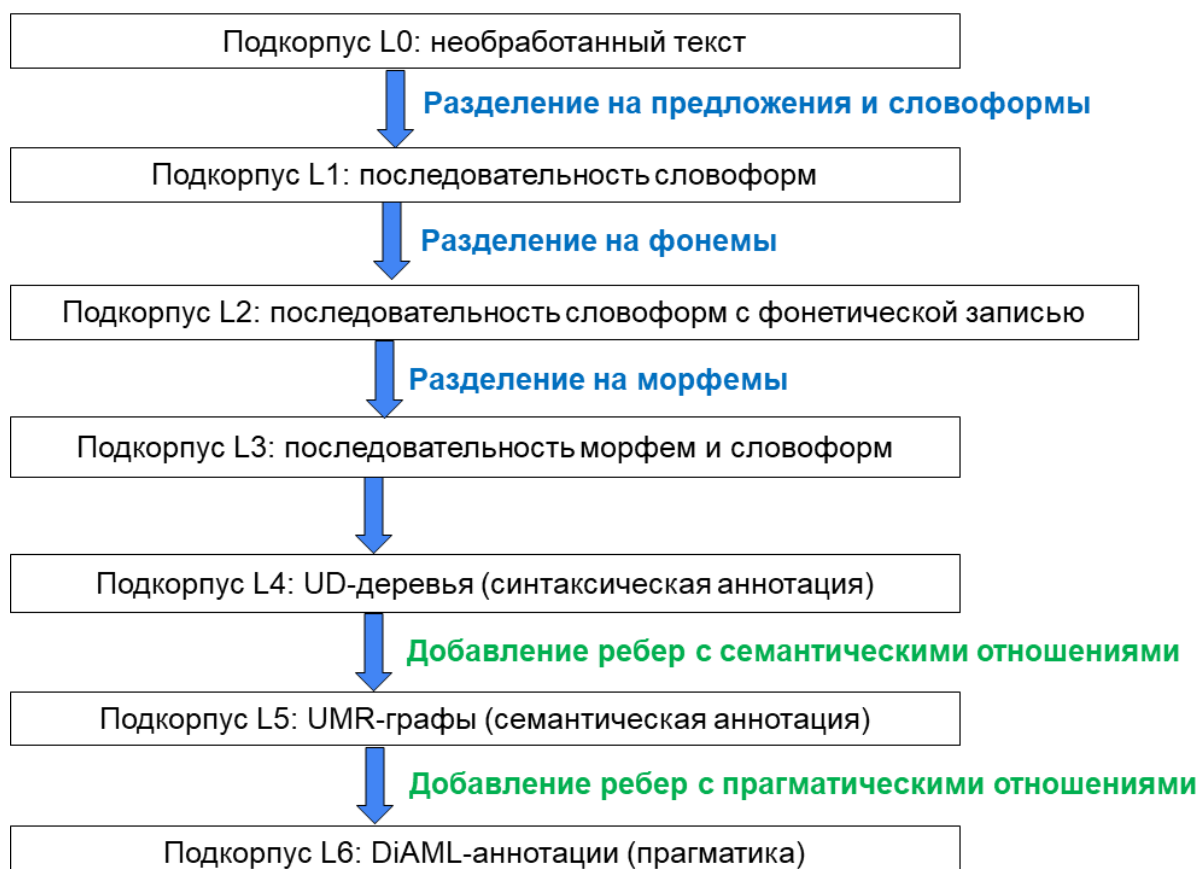


Рис. 2. Конвейер многоуровневой аннотации корпуса.

Связи между эмпирическими данными и лингвистическими единицами обеспечиваются индексными таблицами, поддерживающими двустороннюю навигацию без нарушения однонаправленности ссылок.

2.2. Принцип многоуровневой согласованности

Если трехуровневая архитектура определяет, как организованы данные, то принцип многоуровневой согласованности (семиотическая вертикаль) определяет, какие данные и в каком формате должны быть описаны на каждом уровне. Он является теоретической рамкой, обеспечивающей полноту и непротиворечивость лингвистического описания.

Семиотическая вертикаль представляет собой модель языка как многоуровневой знаковой системы, где каждый уровень соответствует определенному аспекту языковой единицы – от звукового состава до коммуникативной цели высказывания.

Табл. 1. Уровни семиотической вертикали.

Код	Название	Международный стандарт
SL1	Фонологический	IPA
SL2	Морфонологический мост	Продукционные правила
SL3	Морфологический	БНФ, конечные автоматы
SL4	Морфосинтаксический мост	Universal Dependencies
SL5	Синтаксический	Universal Dependencies
SL6	Морфосемантический мост	Словарь соответствий (UD + FrameNet)
SL7	Семантический	UMR, FrameNet, WordNet
SL8	Прагматический	DiAML (ISO 24617-2)

Вертикаль включает восемь уровней (SL1–SL8, см. табл. 1). Каждый уровень опирается на одну из нотаций, принятых в качестве международного стандарта в определенных лингвистических сообществах, что обеспечивает совместимость с внешними инструментами и воспроизводимость аннотаций.

В семиотической вертикали использованы уровни, которые называются семиотическими мостами – SL2, SL4, SL6. Особенность семиотических мостов в том, что они используют не собственные лингвистические единицы, а единицы соседних уровней.

SL2 морфонологический мост – связывает фонетический и морфологический уровни, описывая правила выбора алломорфов в зависимости от фонологического контекста (сингармонизм, ассимиляция).

SL4 морфосинтаксический мост – связывает морфологию и синтаксис, описывая аналитические конструкции и многословные выражения (MWE), которые находятся между словом и свободным синтаксическим сочетанием.

SL6 морфосемантический мост – связывает грамматику и семантику, описывая отображение грамматических признаков в семантические роли (компонент D параметрической модели).

Семиотическая вертикаль реализует два разнонаправленных процесса.

Спецификация (снизу вверх) – нижележащий уровень налагает ограничения на вышележащий. Например, фонологический состав определяет, какие алломорфы допустимы в языке; морфологический уровень определяет, какие комбинации морфем возможны.

Интерпретация (сверху вниз) – вышележащий уровень придает смысл единицам нижележащего. Например, семантический уровень интерпретирует грамматическую форму как семантическую роль участника ситуации; прагматический уровень интерпретирует предложение как речевой акт.

Семиотическая вертикаль выступает в роли системообразующего элемента, определяющего содержание трех основных уровней платформы. В отношении онтологического ядра (уровень 1) она задает требование, согласно которому каждый из восьми уровней должен быть описан на формальном метаязыке, будь то IPA, UD, UMR, DiAML или иной стандарт. Для уровня грамматических описаний (уровень 2) вертикаль предписывает наличие для каждого языка правил, обеспечивающих связь между уровнями: в частности, правил синтаксической аннотации (UD), семантического отображения (UMR) и прагматической интерпретации (DiAML). Наконец, для эмпирического слоя (уровень 3) она устанавливает необходимость многослойной аннотации корпусных данных, охватывающей все уровни вертикали – от фонетической транскрипции (L1) до прагматической разметки (L6).

Таким образом, семиотическая вертикаль обеспечивает сквозную интерпретацию языковых единиц – от акустического сигнала до интенционального акта – и служит теоретическим основанием, на котором строится вся архитектура платформы.

2.3. Принцип морфемоцентричной модели данных

После описания общей архитектуры и принципа многоуровневой согласованности становится понятным, почему базовым элементом хранения выбраны

морфема, а не слово. Это решение продиктовано природой элементарно-комбинаторных языков: именно на уровне морфем сосредоточена основная грамматическая информация, а словоформа представляет собой линейную комбинацию морфем.

Перечислим преимущества морфемоцентричного подхода.

1. Экономичность хранения благодаря перечислимости морфем. В отличие от словоформ и предложений, множество морфем в языке является конечным и перечислимым, его можно полностью описать в виде словаря. Морфемоцентричная модель хранит только этот словарь (несколько тысяч морфем на язык) и правила их сочетания (граф Next, описанный далее), а все словоформы порождаются динамически по запросу. Это соответствует классическому принципу формальных грамматик: конечное множество правил порождает бесконечное множество цепочек.

2. Принцип «одна морфема – одна граммема». В агглютинативных языках каждый аффикс выражает, как правило, одну грамматическую категорию, в отличие от флективных языков, где один аффикс может кодировать несколько признаков одновременно. Это обеспечивает однозначность на морфологическом уровне: каждой морфеме соответствует ровно одна граммема из онтологического ядра. При этом одна граммема может отображаться в разные семантические роли в зависимости от контекста (активированного фрейма), но сам выбор роли детерминирован правилами отображения (компонент D). Таким образом, принцип «одна морфема – одна граммема» гарантирует простоту и предсказуемость морфологического анализа, а многозначность грамматических показателей разрешается на более высоких уровнях (синтаксическом и семантическом) с использованием конечного набора правил.

3. Минимальное количество исключений. В агглютинативных языках правила сочетаемости морфем имеют очень мало исключений: фонологические чередования (сингармонизм, ассимиляция) подчиняются четким и регулярным правилам, а неправильные глаголы или супплетивизм практически отсутствуют. Это позволяет построить детерминированную формальную модель (граф Next), которая покрывает подавляющее большинство словоформ. К примеру, для татарского языка это покрытие составляет более 90%, что отличает эту модель от

модели флективных языков, где многочисленные исключения делают применение подобной модели для флективных языков неэффективной.

4. Единообразное описание синтетических и аналитических форм. В агглютинативных языках одно и то же грамматическое значение может выражаться как аффиксом (синтетически), так и отдельным служебным словом (аналитически). В морфемоцентричной модели и аффиксы, и аналитические морфемы (послоги, частицы, вспомогательные глаголы) являются узлами одного и того же графа Next с грамматической привязкой к граммеме. Различие между ними состоит только в атрибуте узла, определяющем признак написания: слитно или раздельно. Это позволяет обрабатывать синтетические и аналитические конструкции единым механизмом обхода графа, в отличие от лексикоцентричных подходов, где они описываются разными механизмами (морфология vs синтаксис).

5. Морфемоцентричность модели позволяет реализовывать сингармонизм. Сингармонизм (гармония гласных) – это ключевая фонетическая особенность тюркских языков. В морфемоцентричной модели он реализуется на уровне структуры данных: каждый аффикс представлен множеством алломорфов с условиями выбора, причем связь между этими алломорфами осуществляется по правилам сингармонизма. Это гарантирует корректность словоформ без дополнительной логики в коде анализатора.

6. Граф Next как единый механизм фиксации порядка и генерации словоформ. Морфемы связаны отношением Next, которое образует направленный граф допустимых последовательностей. Этот граф одновременно решает две задачи:

- фиксация порядка – явное сохранение последовательности аффиксов, критичное для элементарно-комбинаторных языков, где изменение порядка морфем изменяет смысл словоформы (например, *masadakilerin* против *masaninkilerde*, структура которых была описана ранее);
- детерминированная генерация – обход графа от корня до терминального узла порождает все теоретически возможные словоформы, что особенно важно для малоресурсных языков, где нет больших корпусов для обучения статистических моделей.

Единый формализм для двух задач – фиксации и генерации – является ключевым архитектурным преимуществом.

7. Единообразное и экономичное описание родственных языков благодаря перечислимости морфем и принципу «одна морфема – одна граммема». Морфемы являются значимыми единицами, общими для всей языковой семьи. Сравнение языков упрощается по следующим причинам:

- Перечислимость. Множество морфем в каждом языке конечно, поэтому сравнение сводится к сопоставлению конечных наборов морфем и правил их сочетания, а не к сравнению бесконечных множеств словоформ или предложений.
- Одна морфема – одна граммема. Каждая морфема выражает ровно одну грамматическую категорию, поэтому сравнение ведется на уровне отдельных граммем, а не сложных пучков признаков.
- Одна корневая морфема – один или несколько синсетов WordNet. Лексическая многозначность обрабатывается на уровне семантического анализа, но даже при многозначности сравнение ведется на уровне конечных множеств (корневых морфем и синсетов).
- Одна морфема – одна семантическая роль в UMR. В рамках активированного фрейма каждая морфема, включая аффиксы, выражает ровно одну семантическую роль, что позволяет строить однозначные правила отображения в UMR-графы.

Это позволяет составлять таблицы соответствия граммем и морфем для разных языков, выявлять общетюркские морфемы, переносить знания на малоресурсные языки и параметризовать языки без необходимости описывать все словоформы заново.

8. Естественное представление в виде графа и нативная поддержка СУБД. Морфемоцентричная модель данных имеет прямое структурное соответствие с графовым представлением: морфемы становятся узлами, отношение Next – ребрами, а словоформы – путями в графе. Это позволяет использовать графовые СУБД (Memgraph, Neo4j) для хранения и обработки данных с высокой производительностью:

- обход пути от корня до терминального узла выполняется за $O(k)$ операций, где k – длина словоформы (число аффиксов);

- поиск по морфемной структуре («найти все словоформы, содержащие аффикс X после аффикса Y») реализуется как навигация по ребрам Next;
- семантический поиск с использованием иерархии концептов WordNet выполняется как обход по ребрам отношения hasSynset и таксономическим связям.

В отличие от реляционных СУБД, где аналогичные операции требуют множества субопераций JOIN и рекурсивных запросов, графовая СУБД выполняет их быстрее за счет своей структуры, что особенно важно для больших корпусов объемов порядка миллионов токенов и сложных запросов.

2.4. Принцип параметризации языков

Поскольку тюркские языки близкородственны, их описания имеют общее ядро. Для учета различий и реализации наследования между языками используется параметрический подход, реализованный через четырехкомпонентную систему:

- 1) компонент А – системные параметры (фонемный инвентарь, сингармонизм, число падежей, порядок слотов);
- 2) компонент В – лексические параметры (набор алломорфов морфем, морфотактические правила);
- 3) компонент С – эпистемические параметры (степень изученности языка, достоверность данных);
- 4) компонент D – параметры отображения (соответствие граммем семантическим ролям и прагматическим функциям).

2.5. Взаимодействие базовых принципов

Четыре описанных принципа образуют единую систему, где каждый выполняет свою функцию (табл. 2). Эта архитектура обеспечивает единое формализованное пространство для работы с тюркскими языками, где каждый принцип решает свою задачу, а все вместе они образуют систему, пригодную для автоматической обработки, верификации и коллективного развития.

Табл. 2. Взаимодействие принципов архитектуры платформы “TurkLang”.

Принцип	Основной вопрос	Реализация	Связь с другими принципами
Трехуровневая организация	Как структурированы данные?	Уровни 1 (онтология), 2 (грамматика), 3 (эмпирика) с однонаправленными связями	Определяет архитектуру хранения для семиотической вертикали и морфемоцентричной модели
Многоуровневая согласованность (семиотическая вертикаль)	Какие данные и в каком формате должны быть описаны?	8 уровней (SL1–SL8) с международными стандартами (IPA, UD, UMR, DiAML)	Задаёт содержание уровней 1, 2 и 3; определяет, как морфемы интерпретируются на разных уровнях
Морфемоцентричность	Какой элемент является базовым для хранения?	Морфема как атомарная единица	Опирается на универсалии уровня 1; реализуется через систему отношений морфемы из уровня 2; обеспечивает интерпретацию корпусных данных
Параметризация	Как описывать различия между родственными языками?	Компоненты A–D, наследование $I_{var} = I_{parent} + \Delta I_{var}$	Использует универсалии уровня 1; настраивает грамматику уровня 2 под конкретный язык; определяет правила отображения для UMR

ЗАКЛЮЧЕНИЕ

Представлена архитектура платформы “TurkLang” – системного решения для создания единого цифрового пространства обработки лингвистических данных тюркских языков. Архитектура базируется на четырех взаимосвязанных

принципах: трехуровневой организации базы знаний, многоуровневой согласованности (семиотической вертикали), морфемоцентричной модели данных и параметрическом подходе к описанию родственных языков.

Ключевым решением является переход от лексикоцентрической модели к морфемоцентрической, что обусловлено структурно-функциональными особенностями тюркских языков. Морфемоцентричность обеспечивает экономичность хранения, однозначность морфологического анализа (принцип «одна морфема – одна граммема»), минимальное число исключений, единообразное описание синтетических и аналитических форм, естественную реализацию сингармонизма и сквозную генерацию словоформ через граф Next.

Семиотическая вертикаль (SL1–SL8), опирающаяся на международные стандарты (Universal Dependencies, Uniform Meaning Representation), обеспечивает сквозной анализ от фонемы до речевого акта.

Дальнейшее развитие связано с углублением семантической и прагматической аннотаций, развертыванием RDF-триплетора и масштабированием методологии на другие агглютинативные семьи – финно-угорские, монгольские, тунгусо-маньчжурские и эскимосско-алеутские языки. Предложенная архитектура может служить прототипом для создания лингвистических платформ нового поколения, применимых к широкому кругу агглютинативных языков мира.

Благодарности

Работа выполнена в рамках соглашения №12/2025-ПД от 22.12.2025 о предоставлении гранта научным и научно-педагогическим работникам обособленных структурных подразделений АН РТ с целью стимулирования их к защите докторских диссертаций и выполнению НИР.

СПИСОК ЛИТЕРАТУРЫ

1. Гузев В.Г. О некоторых экзотических особенностях тюркских языков («тюркские чудеса») // Актуальные проблемы мировой политики. 2020. № 10. С. 231–245. <https://doi.org/10.21638/11701/26868318.16>
2. Koskeniemi K. Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. Helsinki: University of Helsinki, 1983. 160 p.

3. *Апресян Ю.Д.* Избранные труды. Том II. Интегральное описание языка и системная лексикография. М.: Школа «Языки русской культуры», 1995. 769 с.
4. *Krauwer S.* The basic language resource kit (BLARK) as the first milestone for the language resources roadmap // Proc. International workshop on speech and computer SPECOM-2003, Moscow, Russia, Oct. 2003. P. 8–15.
5. *Kilgarrieff A., Rychlý P., Smrž P., Tugwell D.* The Sketch Engine // Proceedings of the Eleventh EURALEX International Congress, Lorient, France, Jul. 2004. P. 105–116.
6. *Норманская Ю.В., Борисенко О.Д., Белобородов И.Б., Аветисян А.И.* Программная система LingvoDос и возможности, которые она предлагает для документирования и анализа обско-угорских языков // Докл. РАН. Матем., информ., проц. упр. 2022. № 504. С. 60–82.
<https://doi.org/10.31857/S2686954322030055>
7. *Khanna, T., Washington, J.N., Tyers, F.M et al.* Recent advances in Aper-tium, a free/open-source rule-based machine translation platform for low-resource languages // Machine Translation. 2021. Vol. 35. P. 475–502.
<https://doi.org/10.1007/s10590-021-09260-6>
8. *Bakay Ö., Ergelen Ö., Sarmış E., Yıldırım S., et al.* Turkish WordNet KeNet // Proceedings of the 11th Global Wordnet Conference, Virtual, Jan. 2021. P. 166–174.
<https://doi.org/10.18653/v1/2021.gwc-1.19>
9. *Agostini A., Usmanov T., Khamdamov U., Abdurakhmonova N., et al.* UZ-WORDNET: A Lexical- Semantic Database for the Uzbek Language // Proceedings of the 11th Global Wordnet Conference, Virtual, Jan. 2021. P. 8–19.
<https://doi.org/10.18653/v1/2021.gwc-1.2>
10. *Marsan B., Kara N., Ozcelik M., Arican B. N., et al.* Building the Turkish FrameNet // Proceedings of the 11th Global Wordnet Conference, Virtual, Jan. 2021. P. 118–125. <https://doi.org/10.18653/v1/2021.gwc-1.14>
11. *Rezhake M., Kuerban A.* Design of the Uyghur FrameNet Desktop // Lec-ture Notes on Software Engineering. 2015. Vol. 3. No. 1. P. 53–56.
<https://doi.org/10.7763/LNSE.2015.V3.165>

12. Navigli R., Ponzetto S.P. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network // Artificial Intelligence. 2012. Vol. 193. P. 217–250.

<https://doi.org/10.1016/j.artint.2012.07.001>

13. Gangemi A., Alam M., Asprino L., Presutti V., et al. Framester: A Wide Coverage Linguistic Linked Data Hub. // Knowledge Engineering and Knowledge Management. EKAW 2016. Lecture Notes in Computer Science. 2016. Vol. 10024. P. 239–254. https://doi.org/10.1007/978-3-319-49004-5_16

14. Nivre J., de Marneffe M-C., Ginter F., Hajič J., et al. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection // Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020), Marseille, France, May. 2020. P. 4034–4043.

ARCHITECTURE OF THE LINGUISTIC PLATFORM TURKLANG

A.R. Gatiatullin^[0000-0003-3063-8147]

Tatarstan Academy of Sciences, Kazan, Russia

ayrat.gatiatullin@gmail.com

Abstract

The article describes the architecture of the TurkLang multifunctional linguistic platform, currently being developed at the Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan. The system is based on a three-level knowledge base: linguistic universals, descriptions of units of specific languages, and empirical speech data. The knowledge base is implemented as a graph, where the primary storage unit is the morpheme – the minimal meaningful unit of language (akin to an atomic data element). This design choice is driven by the element-combinatorial nature of Turkic languages: a word form (the concrete realization of a word in text) is a linear chain consisting of a root and sequentially attached affixes – modifiers, each expressing a single grammatical category (case, number, tense, etc.) in a strictly fixed order. The paper examines levels of abstraction, types of relationships between entities, and mechanisms for integration with external sources (an electronic encyclopedia, lexicographic resources, and GIS data). Special attention is paid to the application

of international standards for syntactic (Universal Dependencies), semantic (Universal Meaning Representation), and pragmatic (DiAML) annotation. A comparison with existing platforms (Sketch Engine, LingvoDoc, Apertium) is provided. To date, the platform's knowledge base covers more than 30 languages and dialects, contains over 237,000 root morphemes and about 40,000 compatibility rules, while the corpus component comprises over 180 million of word occurrences.

Keywords: *linguistic platform, knowledge graph, Turkic languages, morpheme-centricity, three-level architecture, Universal Dependencies, Universal Meaning Representation, agglutinative languages, semiotic vertical.*

REFERENCES

1. Guzev V.G. O nekotorykh ekzoticheskikh osobennostyakh tyurkskikh yazykov («tyurkskie chudesa») // Aktual'nye problemy mirovoy politiki. 2020. Vol. 10. P. 231–245. <https://doi.org/10.21638/11701/26868318.16>
2. Koskeniemi K. Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. Helsinki: University of Helsinki, 1983. 160 p.
3. Apresyan Y.D. Izbrannye trudy. Tom II. Integral'noe opisanie yazyka i sistemnaya leksikografiya. Moscow: Shkola “Yazyki russkoi kul'tury”, 1995. 769 p.
4. Krauwer S. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap // Proc. International workshop on speech and computer SPECOM-2003, Moscow, Russia, Oct. 2003. P. 8–15.
5. Kilgariff A., Rychlý P., Smrž P., Tugwell D. The Sketch Engine // Proceedings of the Eleventh EURALEX International Congress, Lorient, France, Jul. 2004. P. 105–116.
6. Normanskaya Y.V., Borisenko O.D., Beloborodov I.B., Avetisyan A.I. The Software System LingvoDoc and the Possibilities It Offers for Documentation and Analysis of Ob-Ugric Languages // Doklady Mathematics. 2022. Vol. 105, No. 3. P. 187–206. <https://doi.org/10.31857/S2686954322030055>
7. Khanna, T., Washington, J.N., Tyers, F.M et al. Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages // Machine Translation. 2021. Vol. 35. P. 475–502. <https://doi.org/10.1007/s10590-021-09260-6>

8. Bakay Ö., Ergelen Ö., Sarmış E., Yıldırım S., et al. Turkish WordNet KeNet // Proceedings of the 11th Global Wordnet Conference, Virtual, Jan. 2021. P. 166–174. <https://doi.org/10.18653/v1/2021.gwc-1.19>
9. Agostini A., Usmanov T., Khamdamov U., Abdurakhmonova N., et al. UZWORDNET: A Lexical- Semantic Database for the Uzbek Language // Proceedings of the 11th Global Wordnet Conference, Virtual, Jan. 2021. P. 8–19. <https://doi.org/10.18653/v1/2021.gwc-1.2>
10. Marsan B., Kara N., Ozcelik M., Arican B. N., et al. Building the Turkish FrameNet // Proceedings of the 11th Global Wordnet Conference, Virtual, Jan. 2021. P. 118–125. <https://doi.org/10.18653/v1/2021.gwc-1.14>
11. Rezhake M., Kuerban A. Design of the Uyghur FrameNet Desktop // Lecture Notes on Software Engineering. 2015. Vol. 3. No. 1. P. 53–56. <https://doi.org/10.7763/LNSE.2015.V3.165>
12. Navigli R., Ponzetto S.P. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network // Artificial Intelligence. 2012. Vol. 193. P. 217–250. <https://doi.org/10.1016/j.artint.2012.07.001>
13. Gangemi A., Alam M., Asprino L., Presutti V., et al. Framester: A Wide Coverage Linguistic Linked Data Hub. // Knowledge Engineering and Knowledge Management. EKAW 2016. Lecture Notes in Computer Science. 2016. Vol. 10024. P. 239–254. https://doi.org/10.1007/978-3-319-49004-5_16
14. Nivre J., de Marneffe M-C., Ginter F., Hajič J., et al. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection // Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020), Marseille, France, May. 2020. P. 4034–4043.

СВЕДЕНИЯ ОБ АВТОРАХ



ГАТИАТУЛЛИН Айрат Рафизович – окончил Казанский государственный университет в 1994 г., к. т. н. (2002). Ведущий научный сотрудник Института прикладной семиотики Академии наук Республики Татарстан. Автор более 100 научных трудов.

Ayrat Rafizovich GATIATULLIN – graduated from Kazan State University in 1994, Candidate of Technical Sciences (2002). Leading researcher at the Institute of Applied Semiotics of Tatarstan Academy of Sciences. List of scientific works includes more than 100 publications.

email: ayrat.gatiatullin@gmail.com

ORCID: 0000-0003-3063-8147

Материал поступил в редакцию 26 июня 2026 года