

УДК 004.822; 81'33

<https://doi.org/10.25587/3034-7378-2026-14-3-5-21>



Оригинальная научная статья

Модель универсального лингвистического графа знаний тюркских языков

А. Р. Гатиатуллин¹  , Н. А. Прокопьев¹ , Н. З. Абдурахмонова² ,
А. А. Касиева³ 

¹Институт прикладной семиотики Академии наук Республики Татарстан,
Казань, Россия

²Национальный университет Узбекистана им. Мирзо Улугбека,
Ташкент, Узбекистан

³Кыргызско-турецкий университет «Манас»,
Бишкек, Кыргызстан

 ayrat.gatiatullin@gmail.com

Аннотация

В статье описывается новая модель универсального лингвистического графа знаний для тюркских языков как основа для создания онтолингвистических ресурсов. В предлагаемой модели знания о мире формализованы в виде онтологического ядра, объединяющего таксономическую и ситуационную онтологии (по аналогии с WordNet и FrameNet) и тесно взаимосвязанного с лингвистическими данными тюркских языков. Актуальность работы обусловлена мало-ресурсностью большинства тюркских языков и потерей эффективности нейросетевых подходов при обработке агглютинативных языков с богатой морфологией, что ограничивает применение больших языковых моделей к тюркским языкам. Цель проекта – разработка новой модели универсального лингвистического графа знаний для тюркских языков. Для реализации этой цели нами решались следующие задачи, такие как объединение знаний о мире и глубоких лингвистических сведений на всех уровнях языковой системы; построение графа по принципу семиотической вертикали; обеспечение морфемцентричности модели; интеграция с международными языковыми стандартами. В качестве основополагающего метода предлагается прагматически-ориентированный подход, фокусирующий предметную область на структурно-функциональных особенностях тюркских языков. Граф знаний строится по принципу

© Гатиатуллин А. Р., Прокопьев Н. А., Абдурахмонова Н. З.,
Касиева А. А., 2026

семиотической вертикали, объединяющей пять уровней – фонологический, морфологический, синтаксический, семантический и прагматический, а также три межуровневых моста (морфонологический, морфосинтаксический и морфосемантический). Ключевая особенность модели – морфемоцентричность, где морфема выступает базовой лингвистической единицей, связанной с единицами всех языковых уровней модели, что принципиально отличает подход от словоцентричных моделей для флективных языков. Модель интегрирована с международными языковыми стандартами: IPA, Universal Dependencies, UDS, UMR, FrameNet, WordNet и DiAML, каждый из которых используется для описания языковых элементов и отношений на своем языковом уровне. Модель уже нашла практическое применение в нескольких программных продуктах, таких как лингвистический портал «Тюркская морфема» и новой версии татарского корпуса «Туган тел». Перспективы включают повышение интерпретируемости LLM, улучшение качества машинного перевода и унификацию аннотаций электронных корпусов.

Ключевые слова: граф знаний, лингвистический граф знаний, тюркские языки, семиотическая вертикаль, онтология, агглютинативные языки, Uniform Meaning Representation, Universal Dependencies, тюркология, лингвистические ресурсы

Для цитирования: Гатиатуллин А.Р., Прокопьев Н.А., Абдурахмонова Н.З., Касиева А.А. Модель универсального лингвистического графа знаний тюркских языков. *Арктика XXI век*. 2026;14(3):5–21. DOI: <https://doi.org/10.25587/3034-7378-2026-14-3-5-21>

Original article

A model of a universal linguistic knowledge graph for Turkic languages

*Ayrat R. Gatiatullin*¹  , *Nikolai A. Prokopyev*¹ ,
*Nilufar Z. Abdurakhmonova*² , *Aida A. Kasieva*³ 

¹Institute of Applied Semiotics of Tatarstan Academy of Sciences,
Kazan, Russia

²National University of Uzbekistan after Mirzo Ulugbek, Tashkent, Uzbekistan

³Kyrgyz-Turkish Manas University, Bishkek, Kyrgyzstan

 ayrat.gatiatullin@gmail.com

Abstract

The article describes a new model of a universal linguistic knowledge graph for Turkic languages as a basis for creating ontolinguistic resources. In the model, world knowledge is formalized as an ontological core that integrates taxonomic

and situational ontologies (by analogy with WordNet and FrameNet) and is closely interrelated with linguistic data from Turkic languages. The relevance of the work is determined by the low-resource nature of most Turkic languages and the loss of effectiveness of neural network approaches in processing agglutinative languages with rich morphology. The goal of the project is to develop a new model of a universal linguistic knowledge graph for Turkic languages. The objectives are: integrating world knowledge and deep linguistic information at all levels of the language system; constructing the graph according to the principle of a semiotic vertical; ensuring the morphemocentricity of the model; and integrating it with international language standards. The foundational method is a pragmatically-oriented approach that focuses the research domain on the structural and functional features of Turkic languages. The knowledge graph is constructed according to the principle of a semiotic vertical that unites five levels – phonological, morphological, syntactic, semantic, and pragmatic – as well as three inter-level bridges (morphonological, morphosyntactic, and morphosemantic). A key feature of the model is its morphemocentricity: the morpheme acts as the basic linguistic unit, connected with units of all language levels of the model, which fundamentally distinguishes this approach from word-centric models for fusional languages. The model is integrated with international standards: IPA, Universal Dependencies, UDS, UMR, FrameNet, WordNet, and DiAML, each of which is used to describe linguistic elements and relations at its own level. The model has already found practical application in several software products, such as the linguistic portal “Turkic Morpheme” and the new version of the Tatar corpus “Tugan Tel”. Prospects include improving the interpretability of LLMs, improving the quality of machine translation, and unifying annotations of electronic corpora.

Keywords: knowledge graph, linguistic knowledge graph, Turkic languages, semiotic vertical, ontology, agglutinative languages, Uniform Meaning Representation, Universal Dependencies, Turkology, linguistic resources

For citation: Gatiatullin A.R., Prokopyev N.A., Abdurakhmonova N.Z., Kasieva A.A. A model of a universal linguistic knowledge graph for Turkic languages. *Arctic XXI Century*. 2026;14(3):5–21 (in Russian). DOI: <https://doi.org/10.25587/3034-7378-2026-14-3-5-21>

Введение

Сегодня практически все ведущие мировые IT-компании участвуют в гонке по разработке больших языковых моделей (Large Language Model – LLM). Если раньше разработка больших языковых моделей входила больше в сферу научных исследований, то на данный момент эти разработки из области научных исследований перешли в область индустриальных разработок. Это связано с тем, что необходимые техно-

логии уже достаточно зрелые и показывают впечатляющие результаты. Однако следует отметить, что для научных исследований еще остаются определенные ниши. В основном это проблемные области, которые не решаются обычным увеличением данных, мощностей процессора и количеством параметров. Какие это проблемные области? Одна из таких проблемных областей – малоресурсность основной массы языков. Следующая проблема – обеспечение достоверности и надежности результатов, выдаваемых LLM. Обе проблемы связаны как с архитектурой самих моделей, которые базируются на аппарате математической статистики, так и с недостоверными, непроверенными данными, которыми изобилуют источники для обучения и из которых черпают информацию все языковые модели. Для малоресурсных языков недостаток достоверной и проверенной информации является еще более острой проблемой.

Как исследователи, так и IT-компании ищут способы повышения надежности, достоверности и достаточности информации. Одним из активно развивающихся направлений является подход с использованием гибридных технологий – нейросимвольный подход, представляющий собой комбинирование символьного и нейросетевого искусственного интеллекта. Если нейросетевой компонент в таких системах извлекает сущности из неструктурированных текстов и структурирует данные, то символьный ИИ, как правило, используется для определения достоверных данных и проверки результатов, выдаваемых нейросетевыми модулями.

В данной статье предлагается решение для тюркских языков в виде модели универсального лингвистического графа знаний, служащей основой для создания онтолингвистических ресурсов и символьного ИИ.

Проблемы и существующие подходы

На сегодняшний день наиболее популярными инструментами, используемыми в гибридных технологиях с нейронными сетями, являются онтологии и графы знаний. Самые распространенные – это RAG-технологии (Retrieval-Augmented Generation), которые представляют собой предварительный поиск релевантной информации из внешних баз данных [1]. При наличии достоверных баз данных этот подход позволяет снижать количество «галлюцинаций». GraphRAG [2, 3] – это графовая разновидность RAG-подхода. Данный подход осуществляет поиск в онтологической базе данных и улучшает качество ответа за счет логических и семантических взаимосвязей.

Особую значимость нейросимвольный подход приобретает в контексте малоресурсных языков – языков, для которых отсутствуют большие массивы текстов, необходимые для обучения мощных LLM. Для таких языков экстенсивный путь индустрии оказывается тупиковым:

невозможно нарастить объем данных там, где их попросту нет. Именно здесь гибридные технологии становятся не просто альтернативой, а единственным реалистичным путем создания качественных интеллектуальных систем. Лингвистические онтологии и графы знаний, разработанные лингвистами вручную, используются для компенсации нехватки эмпирических данных. Символьный модуль системы обеспечивает грамматическую корректность и семантическую точность, используя структурированные знания.

Для тюркских языков, с одной стороны, актуальна проблема малоресурсности, а с другой стороны, возникают проблемы потери эффективности работы нейронных сетей, связанные со структурно-функциональными особенностями этих языков [4, 5]. Тюркские языки относятся к агглютинативному типу с богатой морфологией и целым рядом дополнительных структурно-функциональных особенностей [6]. Из-за этих особенностей чистые нейросетевые модели теряют свою эффективность, которую они показывают на индоевропейских языках. Отсутствие в нейросетевых архитектурах учета формальных морфологических или синтаксических особенностей тюркских языков создает определенные трудности для работы больших языковых моделей. Особенно это проявляется на редких словоформах или словоформах, полностью отсутствующих в корпусах, на которых производится обучение моделей.

Среди наиболее известных работ по реализации лингвистических баз знаний следует отметить проект OntoLex-Lemon, который представляет собой RDF-словарь для представления лексической информации в связанных данных [7]. OntoLex-Lemon стал де-факто стандартом для моделирования лексиконов в RDF и был использован при реализации многих известных лингвистических баз знаний, таких как Princeton WordNet, DBnary, FrameNet и VerbNet. OntoLex-Lemon напрямую развивает идеи LMF (Lexical Markup Framework) [8], но делает это в RDF-парадигме, обеспечивая совместимость с принципами связанных данных. Модульная архитектура OntoLex-Lemon обеспечивает не только гибкость при выборе необходимых компонентов, но и возможность расширения модели под конкретные задачи. Однако OntoLex-Lemon является продуктом, центральным элементом которого являются лексемы, в связи с чем он теряет свою эффективность при решении ряда задач по формальному описанию и обработке тюркских языков. В моделях и базах знаний, работающих с агглютинативными языками, более продуктивными являются морфемоцентричные модели. Это связано с тем, что такие свойства, как регулярность и предсказуемость, а также наличие четких границ между морфемами делают тюркские языки идеальным

объектом для применения графовых, онтологических и автоматных технологий, в которых узлами графов выступают морфемы.

Эта особенность и послужила источником идеи описания всех тюркских языков в виде единой многоуровневой системы лингвистических моделей на основе графов знаний. Эту модель в виде универсального лингвистического (семиотического) графа знаний, которая описывает тюркские языки на пяти языковых уровнях и имеет несколько межуровневых мостов, мы назвали графом знаний TurkLang [9]. В этой статье описывается дальнейшее развитие модели описания тюркских языков в виде семиотической вертикали, которая состоит из пяти языковых уровней и трех межуровневых мостов, что позволяет наиболее полно отразить типологические особенности тюркских языков с точки зрения их компьютерной обработки.

Методология

Узлами графа являются лингвистические единицы разных языковых уровней, а ребрами – отношения между этими лингвистическими единицами. Все лингвистические единицы модели поделены на базовые, конструируемые и полуконструируемые. Отличительной особенностью базовых единиц является то, что их количество конечно, и они определяются в словарях языка. Конструируемые единицы, соответственно, образуются из базовых единиц. Особенными являются многословные единицы, которые также необходимо хранить в языковых словарях, поскольку значения этих единиц не выводимы из значений их компонентов – они относятся к полуконструируемым единицам.

Основным описательным элементом языка в графе является морфема и, соответственно, модель является морфемоцентричной. Этот подход принципиально отличается от моделей, созданных для флективных языков (например, индоевропейских), в которых в качестве основного элемента используется слово, а модели являются словоцентричными. Морфемы в модели графа знаний TurkLang подразделяются на корневые, аффиксальные и аналитические и могут выступать как лексические морфемы, и как грамматические, от чего зависит система отношений каждой морфемы. Так, например, лексическая морфема вступает в систему отношений с семантическими концептами, а грамматическая морфема – с грамматическими значениями. Для агглютинативных тюркских языков именно морфема является тем естественным кирпичиком, который несет грамматическое и семантическое значение, слово же представляет собой линейную цепочку морфем, собранных по определенным правилам.

Граф знаний TurkLang мы называем семиотическим, поскольку этот граф объединяет в себе грамматические, семантические и праг-

матические свойства лингвистических единиц в единой системе, построенной по принципу семиотической вертикали (рис. 1).

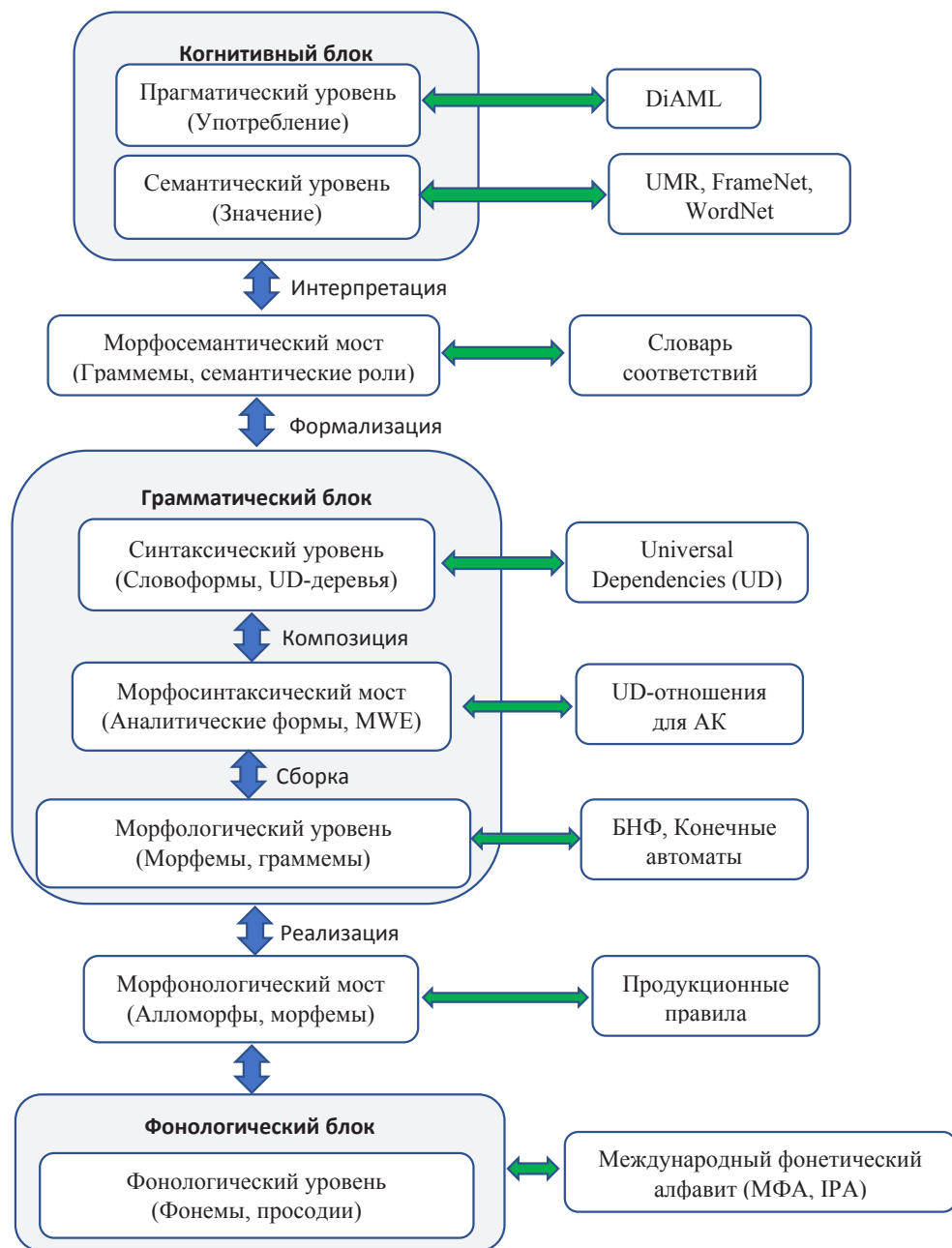


Рис. 1. Структура семиотической вертикали

Fig. 1. Structure of the semiotic vertical

Семиотической вертикалью называется иерархическая модель, описывающая язык как знаковую систему и позволяющая проследить путь языковой единицы от ее фонетической реализации до коммуникативной функции в высказывании. Вертикаль объединяет три блока – фонологический, грамматический и когнитивный – и охватывает пять уровней и три межуровневых моста языковой системы. Разница между уровнями и мостами в том, что уровни описывают свои языковые единицы, а межуровневые мосты оперируют с языковыми единицами соседних уровней.

Одна из целей, заложенная при разработке вертикали, заключается в интеграции в нее международных стандартов, чтобы ресурсы на ее основе были доступны для анализа с помощью существующего программного инструментария. Поэтому на каждом уровне выделены соответствующие стандарты, с которыми он интегрирован.

Далее рассмотрим блоки, уровни и межуровневые мосты.

Фонологический блок состоит из одного уровня – фонологического, который описывает материальную субстанцию языка – инвентарь фонем и просодические характеристики, используя международный фонетический алфавит (МФА).

Морфофонологический мост связывает между собой фонологический и морфологический уровни, определяя выбор алломорфов морфем в зависимости от фонологического контекста.

Грамматический блок включает морфологический и синтаксический уровни, взаимосвязанные между собой морфосинтаксическим мостом.

Морфологический уровень описывает структуру слова: базовый словарь морфем с правилами их сочетаемости в рамках словоформы. Для формального описания грамматических значений морфем в модели представлены словари грамем и грамматических категорий. Как такового единого стандарта этот уровень не имеет, связываясь с существующими описательными механизмами формальных грамматик (к примеру, формы Бэкуса-Наура) и конечных автоматов (таких как HFST [10]).

Синтаксический уровень описывает структуру предложения в виде типов синтаксических единиц и правил их взаимосвязи между собой. В отличие от фонологического и морфофонологического уровней здесь приводится не базовый словарь лингвистических единиц этого уровня, а набор типов, поскольку единицы этого уровня конструируемые и они описываются через набор свойств. В качестве формального аппарата описания синтаксического уровня используются деревья синтаксических зависимостей, а в качестве международного стандарта описания синтаксических единиц – стандарт Universal Dependencies (UD) [11].

Морфосинтаксический мост описывает аналитические морфемы в виде послелогов и частиц, а также аналитические формы и аналитические конструкции. Конструкции этого уровня являются промежуточными между синтаксисом и морфологией, причем для разных тюркских языков эти конструкции отличаются. Например, если в турецком языке вопросительная частица пишется отдельно и является аналитической морфемой, то в татарском языке вопросительная частица пишется слитно и входит в раздел аффиксальных морфем. Кроме того, на данном уровне описаны многословные выражения – это конструкции, представляющие собой синтаксическое сочетание, значение которого не выводимо из значения его составляющих. Морфосинтаксический мост связывает их значения либо со статическими концептами, либо с типовыми ситуациями.

Морфосемантический мост обеспечивает связь между грамматическими и семантическими значениями. Грамматические значения представляют граммы и грамматические категории, а семантические значения – это семантические роли в ситуационно-фреймовых моделях.

Когнитивный блок содержит два уровня: семантический и прагматический.

Семантический уровень содержит механизмы для формального представления смысла; это словари для описания семантики типа WordNet [12] и FrameNet [13]. Если словарь WordNet содержит иерархическую систему для описания статической семантики корневых морфем, то FrameNet представляет собой словарь типовых ситуаций и используется для описания динамической семантики. В качестве международного стандарта применяется Uniform Meaning Representation (UMR) [14] – язык описания семантики текстов, который является дальнейшим развитием стандарта Abstract Meaning Representation (AMR) [15]. Этот язык позволяет описывать смысл текста в виде семантических графов со стандартными типами понятий и отношений между ними.

Прагматический уровень – это уровень, который описывает интерпретации высказывания в виде коммуникативных ситуаций, речевых актов и диалоговых зависимостей. На данный момент это наименее развитый уровень в научных исследованиях, основным стандартом для него является язык DiAML [16].

Уровни в семиотической вертикали оказывают взаимное влияние друг на друга. Нижележащие уровни накладывают формальные ограничения на вышележащие уровни, а вышележащие уровни придают дополнительные значения нижним уровням. Благодаря этому свойству модель может быть использована как для процесса анализа, так и для осуществления синтеза речевых фрагментов.

Результаты и обсуждение

В настоящее время модель TurkLang является не только чисто теоретической моделью, она уже нашла практическое воплощение в ряде программных продуктов и баз знаний. Основным продуктом, созданным на ее основе, является лингвистический портал «Тюркская морфема» [17]. Вторая разработка – это новая версия электронного корпуса татарского языка «Туган тел» [18], которая интегрируется с порталом «Тюркская морфема» в единую распределенную систему. Сегодня новая версия электронного корпуса находится в стадии разработки и пока не доступна в общем доступе.

Модель TurkLang обеспечивает совместимость на разных языковых уровнях и взаимодополняемость лингвистических ресурсов для разных тюркских языков. По мнению авторов, в перспективе эта модель могла бы стать основой для построения единого онто-лингвистического пространства тюркских языков. Благодаря этой модели достигается унификация цифрового представления тюркских лингвистических единиц и отношений между ними.

В рамках архитектурного решения, заложенного в графе знаний TurkLang, принципиально важным представляется вопрос о формальном аппарате, обеспечивающем переход от грамматической структуры предложения к его глубинному смыслу. В предлагаемой модели этот переход реализуется не через единый стандарт, а через каскадное применение трех взаимосвязанных спецификаций, каждая из которых отвечает за свой этап семантико-синтаксической трансформации, а именно Universal Dependencies (UD), Universal Decompositional Semantics (UDS) и Uniform Meaning Representation (UMR). При этом, с одной стороны, необходимо было решить задачу адаптации этих стандартов под структурно-функциональные особенности тюркских языков, а с другой стороны, обеспечить взаимосвязь между самими этими стандартами для обеспечения полного комплексного описания тюркских языков.

Отправной точкой здесь выступает синтаксический уровень, для описания которого задействован стандарт UD. Суть его применения сводится к тому, что любое предложение получает формальное представление в виде дерева зависимостей, где словоформы выступают вершинами, а дуги между ними маркируют типы подчинительной связи – подлежащее, дополнение, обстоятельство и т. д. Важно подчеркнуть, что система из 37 универсальных отношений, предлагаемая UD, является достаточной для единообразной аннотации не только английских или русских языков, но и для языков со сложной морфосинтаксикой, таких как тюркские. Для последних данный стандарт обогащается так

называемыми расширенными зависимостями (Enhanced Dependencies), которые специально предназначены для фиксации специфических оборотов, включая пассивные конструкции с агентивным дополнением или относительные придаточные, характерные для агглютинативного строя.

UD фиксирует лишь поверхностную структуру, но не отвечает на вопрос о том, кто именно совершил действие и с какой степенью реальности оно произошло. Эту функцию берет на себя морфосемантический мост, реализованный через стандарт UDS [19]. В отличие от предыдущего уровня, UDS не существует сам по себе – он надстраивается непосредственно над деревьями зависимостей, полученными из UD. Если описывать этот процесс поэтапно, то сначала осуществляется парсинг, затем на основе синтаксиса извлекается предикатно-аргументная сетка, и, наконец, эта сетка покрывается уникальным слоем – скалярными атрибутами. Именно эти атрибуты (фактуальность, генеричность или градуированные прото-роли) являются особенностью UDS, поскольку они позволяют уйти от жестких бинарных классификаций. Для тюркского словообразования, где аффиксы часто передают не просто время или лицо, но и тонкие оттенки достоверности или интенсивности, данный инструмент является практически незаменимым.

Можно сказать, что вершиной семиотической вертикали на данный момент является семантический уровень, специфицируемый через UMR. В данном стандарте осуществляется финальная сборка смысла, и здесь уже недостаточно анализа отдельного предложения. UMR интересен своей двухуровневой организацией: сначала строится граф для отдельного предложения (предикаты, аргументы, именованные сущности), а затем формируется граф для всего документа в целом. Второй уровень крайне важен, так как позволяет отслеживать кореферентность и временные связи на протяжении всего текста, что невозможно сделать изолированно. Стоит отметить, что UMR создавался специально с учетом типологического разнообразия языков мира, поэтому он хорошо справляется с агглютинативными языками, даже если для них нет больших размеченных корпусов.

Говоря о связанности этих трех уровней, нельзя не отметить, что между ними существует жесткая логическая преемственность, подкрепленная вычислительными методами. В частности, активно применяются методики «bootstrapping», при которых, имея качественные деревья зависимостей в формате UD, можно автоматически сгенерировать начальную, черновую версию UMR-графа без экспертного вмешательства. Хотя такой подход и не дает стопроцентной точности, для малоресурсных языков он является единственным экономически целесообразным

способом получения структурированных семантических данных. В конечном итоге, предложенная связка стандартов $UD \rightarrow UDS \rightarrow UMR$ создает непрерывный аналитический пайплайн, где синтаксический каркас постепенно обрastaет сначала градуальными семантическими признаками, а затем превращается в полноценную онтологическую структуру с учетом дискурсивных связей, что и является необходимым условием для построения объяснимых нейросимволических систем ИИ в тюркоязычной среде.

Заключение

Разработанная модель универсального лингвистического графа знаний TurkLang, основанная на прагматически-ориентированном подходе применительно к структурно-функциональным особенностям тюркских языков и построенная по принципу семиотической вертикали, объединяет фонологический, грамматический и когнитивный блоки, охватывая пять уровней языковой системы – от фонем до прагматических интерпретаций – и три межуровневых моста. Ключевой особенностью модели является ее морфемоцентричность, когда морфема выступает базовой единицей, через которую устанавливаются связи с единицами всех остальных уровней. Эта особенность принципиально отличает данный подход от словоцентричных моделей, разработанных для флективных языков.

Данная модель не окончательная и является открытой. Уточнение модели происходит в ходе практической реализации лингвистических ресурсов на ее базе и исследовании новых языковых особенностей. Интеграция международных стандартов, описанных в статье, обеспечивает совместимость создаваемого ресурса с существующими инструментами и корпусами. Несмотря на то, что модель уже получила практическую реализацию в виде таких лингвистических ресурсов, как портал «Тюркская морфема» и новая версия электронного корпуса татарского языка «Туган тел», еще остаются открытыми вопросы интеграции в модель прагматического и дискурсивного уровней.

Дальнейшие направления исследований включают гибридные и автоматические методы извлечения таксономических отношений для создания онтологии, а также экспериментальную верификацию модели и оценку прикладных эффектов ее применения. Отдельное направление расширения модели – реализация пространственно-временной культурно-специфичной информации с привлечением геоинформационных систем и методов нечеткой логики. В перспективе представленная модель может стать основой для создания распределенной онто-лингвистической платформы, которая будет объединять несколько лингвистических разработок Института прикладной семиотики АН РТ.

Литература

1. Lewis P., Perez E., Piktus A. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*. 2020:9459–9474. DOI: <https://doi.org/10.5555/3495724.3496517> (на англ.).
2. Edge D., Trinh H., Cheng N. et al. From local to global: a graph RAG approach to query-focused summarization. *arXiv cs.CL 2404.16130*. 2025. DOI: <https://doi.org/10.48550/arXiv.2404.16130> (на англ.).
3. Ломов П.А. Использование онтологий для контекстуализации запросов к большим языковым моделям. *Онтология проектирования*. 2025;15(2):239–248. DOI: <https://doi.org/10.18287/2223-9537-2025-15-2-239-248>
4. Arnett C., Bergen B. Why do language models perform worse for morphologically complex languages? *Proceedings of the 31st International Conference on Computational Linguistics*. UAE. 2025:6607–6623 (на англ.).
5. Yu S., Kulkarni N., Lee H., and Kim J. Syllable-level Neural Language Model for agglutinative language. *Proceedings of the First Workshop on Subword and Character Level Models in NLP*. Copenhagen. 2017:92–96 (на англ.).
6. Гузев В.Г. О некоторых экзотических особенностях тюркских языков («тюркские чудеса»). *Актуальные проблемы мировой политики*. 2020;(10):231–245.
7. McCrae J., Spohr D., Cimiano P. Linking lexical resources and ontologies on the Semantic Web with Lemon. *The Semantic Web: Research and Applications. ESWC 2011. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer. 2011;(6643):245–259. DOI: https://doi.org/10.1007/978-3-642-21034-1_17 (на англ.).
8. Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., and Soria C. Lexical Markup Framework (LMF). *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Italy. 2006 (на англ.).
9. Гатиатуллин А.Р., Прокопьев Н.А., Сулейманов Д.Ш. Модель лингвистических графов знаний тюркских языков. *Онтология проектирования*. 2024;14(3):366–378.
10. Linden K., Axelson E., Drobac S. et al. HFST – A system for creating NLP tools. *Systems and Frameworks for Computational Morphology. SFCM 2013. Communications in Computer and Information Science*. 2013;(380):53–71. DOI: https://doi.org/10.1007/978-3-642-40486-3_4 (на англ.).
11. Marneffe M., Manning C., Nivre J. et al. Universal Dependencies. *Computational Linguistics*. 2021;47(2):255–308. DOI: https://doi.org/10.1162/coli_a_00402 (на англ.).
12. Fellbaum C. *WordNet: An electronic lexical database*. Cambridge. MA: MIT Press. 1998. DOI: <https://doi.org/10.7551/mitpress/7287.001.0001> (на англ.).

13. Baker C.F., Fillmore C.J., Lowe J.B. The Berkeley FrameNet project. *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. 1998;(1):86–90. DOI: <https://doi.org/10.3115/980845.980860> (на англ.).
14. Van Gysel J.E.L., Vigus M., Chun J. et al. Designing a Uniform Meaning Representation for Natural Language Processing. *Künstliche Intelligenz*. 2021;(35):343–360. DOI: <https://doi.org/10.1007/s13218-021-00722-w> (на англ.).
15. Banarescu L., Bonial C., Cai S. et al. Abstract Meaning Representation for Sembanking. *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*. 2013:178–186 (на англ.).
16. Bunt H., Alexandersson J., Choe J.W. et al. ISO 24617-2: A semantically-based standard for dialogue annotation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. 2012:430–437 (на англ.).
17. Gatiatullin A., Suleymanov D., Prokopyev N., Khakimov B. About “Turkic Morpheme” portal. *CEUR Workshop Proceedings*. 2020;(2780):226–243 (на англ.).
18. Mukhamedshin D., Gatiatullin A., Gilmullin R. The new version of the corpus data management system “Tugan Tel” using graph knowledge base. *IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE)*. 2024:1800–1804. DOI: <https://doi.org/10.1109/PIERE62470.2024.10804932> (на англ.).
19. White A.S., Reisinger D., Sakaguchi K. et al. Universal Decompositional Semantics on Universal Dependencies. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016:1713–1723. DOI: <https://doi.org/10.18653/v1/D16-1177> (на англ.).

References

1. Lewis P, Perez E, Piktus A et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*. 2020:9459–9474. DOI: <https://doi.org/10.5555/3495724.3496517>
2. Edge D, Trinh H, Cheng N et al. From local to global: a graph RAG approach to query-focused summarization. *arXiv cs.CL 2404.16130*. 2025. DOI: <https://doi.org/10.48550/arXiv.2404.16130>
3. Lomov PA. Using ontologies to contextualize queries to Large Language Models. *Ontology of Designing*. 2025;15(2):239–248. DOI: <https://doi.org/10.18287/2223-9537-2025-15-2-239-248> (in Russian).
4. Arnett C, Bergen B. Why do language models perform worse for morphologically complex languages? *Proceedings of the 31st International Conference on Computational Linguistics*. UAE. 2025:6607–6623.

5. Yu S, Kulkarni N, Lee H, and Kim J. Syllable-level Neural Language Model for agglutinative language. *Proceedings of the First Workshop on Subword and Character Level Models in NLP*. Copenhagen. 2017:92–96.
6. Guzev VG. About some exotic features of Turkic languages (“Turkic miracles”). *Digest of World Politics*. 2020;(10):231–245 (in Russian).
7. McCrae J, Spohr D, Cimiano P. Linking lexical resources and ontologies on the Semantic Web with Lemon. *The Semantic Web: Research and Applications. ESWC 2011. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer. 2011;(6643):245–259. DOI: https://doi.org/10.1007/978-3-642-21034-1_17
8. Francopoulo G, George M, Calzolari N, Monachini M, Bel N, Pet M, and Soria C. Lexical Markup Framework (LMF). *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Italy. 2006.
9. Gatiatullin AR, Prokopyev NA, Suleymanov DSh. Model of linguistic knowledge graphs of Turkic languages. *Ontologiya proyektirovaniya*. 2024;14(3):366–378 (in Russian).
10. Linden K, Axelson E, Drobac S et al. HFST – A system for creating NLP tools. *Systems and Frameworks for Computational Morphology. SFCM 2013. Communications in Computer and Information Science*. 2013;(380):53–71. DOI: https://doi.org/10.1007/978-3-642-40486-3_4
11. Marneffe M, Manning C, Nivre J et al. Universal Dependencies. *Computational Linguistics*. 2021;47(2):255–308. DOI: https://doi.org/10.1162/coli_a_00402
12. Fellbaum C. *WordNet: an electronic lexical database*. Cambridge. MA: MIT Press. 1998. DOI: <https://doi.org/10.7551/mitpress/7287.001.0001>
13. Baker CF, Fillmore CJ, Lowe JB. The Berkeley FrameNet project. *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. 1998;(1):86–90. DOI: <https://doi.org/10.3115/980845.980860>
14. Van Gysel JEL, Vigus M, Chun J et al. Designing a Uniform Meaning Representation for Natural Language Processing. *Künstliche Intelligenz*. 2021;(35):343–360. DOI: <https://doi.org/10.1007/s13218-021-00722-w>
15. Banarescu L, Bonial C, Cai S et al. Abstract Meaning Representation for Sembanking. *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*. 2013:178–186.
16. Bunt H, Alexandersson J, Choe JW et al. ISO 24617-2: A semantically-based standard for dialogue annotation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. 2012:430–437.
17. Gatiatullin A, Suleymanov D, Prokopyev N, Khakimov B. About “Turkic Morpheme” portal. *CEUR Workshop Proceedings*. 2020;(2780):226–243.
18. Mukhamedshin D, Gatiatullin A, Gilmullin R. The new version of the corpus data management system “Tugan Tel” using graph knowledge base.

IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE). 2024:1800–1804. DOI: <https://doi.org/10.1109/PIERE62470.2024.10804932>

19. White AS, Reisinger D, Sakaguchi K et al. Universal Decompositional Semantics on Universal Dependencies. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016:1713–1723. DOI: <https://doi.org/10.18653/v1/D16-1177>

Об авторах

ГАТИАТУЛЛИН Айрат Рафизович – кандидат технических наук, ведущий научный сотрудник, Институт прикладной семиотики Академии наук Республики Татарстан, г. Казань, Россия, ORCID: <https://orcid.org/0000-0003-3063-8147>, WoS ResearcherID: ABF-8777-2020, Scopus Author ID: 56500678000, Elibrary AuthorID: 161758, SPIN: 4633-2846, e-mail: ayrat.gatiatullin@gmail.com

ПРОКОПЬЕВ Николай Аркадиевич – кандидат технических наук, научный сотрудник, Институт прикладной семиотики Академии наук Республики Татарстан, г. Казань, Россия, ORCID: <https://orcid.org/0000-0003-0066-7465>, WoS ResearcherID: S-3829-2016, Scopus Author ID: 57206891311, Elibrary AuthorID: 999214, SPIN: 2430-9028, e-mail: nikolai.prokopyev@gmail.com

АБДУРАХМОНОВА Нилуфар Зайнобидиновна – доктор филологических наук, профессор, заведующая кафедрой компьютерной лингвистики и прикладного языкознания, Факультет журналистики и узбекской филологии, Национальный университет Узбекистана им. Мирзо Улугбека, г. Ташкент, Узбекистан, ORCID: <https://orcid.org/0000-0001-9195-5723>, WoS ResearcherID: ABC-5275-2021, Scopus Author ID: 57221106516, e-mail: n.abduraxmonova@nuu.uz

КАСИЕВА Аида Аскарбековна – кандидат филологических наук, доцент, профессор, заведующая отделением перевода и переводоведения, Филологический факультет, Кыргызско-Турецкий университет «Манас», г. Бишкек, Кыргызстан, ORCID: <https://orcid.org/0000-0002-9605-7460>, WoS ResearcherID: HGC-3155-2022, Scopus Author ID: 57189928164, e-mail: aida.kasieva@manas.edu.kg

About the authors

Ayrat R. GATIATULLIN – Cand. Sci. (Engineering), Leading Researcher, Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan, Tatarstan, Russia, ORCID: <https://orcid.org/0000-0003-3063-8147>, WoS ResearcherID: ABF-8777-2020, Scopus Author ID: 56500678000, Elibrary AuthorID: 161758, SPIN: 4633-2846, e-mail: ayrat.gatiatullin@gmail.com

Nikolai A. PROKOPYEV – Cand. Sci. (Engineering), Researcher, Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan, Tatarstan, Russia,

ORCID: <https://orcid.org/0000-0003-0066-7465>, WoS ResearcherID: S-3829-2016, Scopus Author ID: 57206891311, Elibrary AuthorID: 999214, SPIN: 2430-9028, e-mail: nikolai.prokopyev@gmail.com

Nilufar Z. ABDURAKHMONOVA – Dr. Sci. (Philology), Professor, Head of the Department of Computational and Applied Linguistics, Faculty of Journalism and Uzbek Philology, National University of Uzbekistan Named after Mirzo Ulugbek, Tashkent, Uzbekistan, ORCID: <https://orcid.org/0000-0001-9195-5723>, WoS ResearcherID: ABC-5275-2021, Scopus Author ID: 57221106516, e-mail: n.abduraxmonova@nuu.uz

Aida A. KASIEVA – Cand. Sci. (Philology), Associate Professor, Professor and Head of the Department of Translations, Faculty of Humanities, Kyrgyz-Turkish Manas University, Bishkek, Kyrgyzstan, ORCID: <https://orcid.org/0000-0002-9605-7460>, WoS ResearcherID: HGC-3155-2022, Scopus Author ID: 57189928164, e-mail: aida.kasieva@manas.edu.kg

Вклад авторов

Гатиатуллин А. Р. – разработка концепции, методологии, руководство разработкой, верификация

Прокопьев Н. А. – подготовка структур баз знаний и программная реализация

Абдурахмонова Н. З. – верификация модели на базе узбекского языка

Касиева А. А. – верификация модели на базе кыргызского языка

Authors' contribution

Ayrat R. Gatiatullin – development of the concept and methodology, management of the development process, verification

Nikolai A. Prokopyev – preparation of knowledge base structures and software implementation

Nilufar Z. Abdurakhmonova – model verification on the Uzbek language

Aida A. Kasieva – model verification on the Kyrgyz language

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов

Conflict of interests

The authors declare no relevant conflict of interests

Поступила в редакцию / Submitted 26.07.2026

Поступила после рецензирования / Revised 19.08.2026

Принята к публикации / Accepted 07.09.2026