УДК 004.822 *Научная статья* DOI: 10.18287/2223-9537-2024-14-3-366-378

Модель лингвистических графов знаний тюркских языков

© 2024, А.Р. Гатиатуллин✉, Н.А. Прокопьев, Д.Ш. Сулейманов

Академия наук Республики Татарстан, Институт прикладной семиотики, Казань, Россия

Аннотация

Описана модель лингвистического графа знаний тюркских языков *TurkLang*, которая положена в основу программных продуктов для компьютерной обработки тюркских языков. Базовыми элементами новой модели лингвистических графов знаний являются минимальные значимые единицы языка – морфемы. В структуре графа знаний отражены свойства морфемы на морфонологическом, морфологическом, синтаксическом и семантическом уровнях. Подобная модель в наибольшей степени соответствует структурно-функциональным особенностям тюркских языков, как языков агглютинативного типа, и позволяет полно и прагматически-ориентированно описывать потенциальные возможности тюркских языков и их проявления в текстах. Свойства модели использованы в программных продуктах, связанных с семантической обработкой текстов, в составе лингвистического портала «Тюркская морфема» и новой версии электронного корпуса татарского языка «Туган тел». Единая модель лингвистического графа знаний тюркских языков, представленная в статье, позволяет обеспечить полную совместимость программных продуктов, реализуемых для тюркских языков, использовать единую систему понятий и терминов в лингвистических исследованиях. Для тюркских языков это актуально, поскольку многие разработчики используют модели, созданные для языков с другой структурой (английской, русской и др.), а эти модели не соответствуют в полной мере структуре тюркских языков, не позволяют отразить весь коммуникативный и когнитивный потенциал и лексико-грамматические особенности тюркских языков.

Ключевые слова: *граф знаний, Интернет-портал, лингвистическая единица, морфема, тюркский язык.*

Цитирование: *Гатиатуллин А.Р., Прокопьев Н.А., Сулейманов Д.Ш.* Модель лингвистических графов знаний тюркских языков // Онтология проектирования. 2024. Т.14, №3(53). С.366-378. DOI: 10.18287/2223-9537-2024-14-3-366-378.

Финансирование: Работа выполнена при поддержке Российского научного фонда (проект 24-21-00453).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

Для решения задач компьютерной обработки языка требуется наличие лингвистических баз данных (БД) и баз знаний (БЗ). По наличию таких ресурсов происходит классификация на богатые лингвистическими ресурсами и малоресурсные языки. Малоресурсные языки – это естественные языки (ЕЯ), для которых не хватает электронных лингвистических ресурсов для обработки языка и речи, в т.ч. одноязычных корпусов, электронных словарей разного типа, орфографических и фонетических транскрипций речи и т.д. [1]. Разница между богатыми лингвистическими ресурсами и малоресурсными языками постоянно нарастает по объективным причинам, в т.ч. из-за структурной разницы языков. Программное обеспечение, разработанное для языков индоевропейского семейства, не всегда применимо для тюркских языков (ТЯ). Компьютерные лингвистические модели, разработанные для индоевропейских языков, не отображают всю полноту структурно-функциональных особенностей ТЯ.

Многолетний опыт в области разработок и использования инфокоммуникационных технологий показывает, что современные средства накопления и обработки знаний неэффектив-

ны и практически не справляются с такими задачами, как поиск и отбор информации в распределённых БД, извлечение знаний, семантический анализ текстовой информации, прежде всего потому, что они не интеллектуальны изначально. И причиной их неинтеллектуальности является, главным образом то, что создаются они с использованием языков программирования, практически представляющих собой подмножество флективно-аналитических языков или искусственных структур, созданных на основе ЕЯ, морфо-синтаксические структуры которых больше ориентированы на реализацию коммуникативных функций, нежели когнитивных.

Основные отличительные особенности ТЯ [2,3]: агглютинативность, сингармонизм, отсутствие грамматического выделения единственного числа и категории рода; редко встречаются исключения из правил, большинство агглютинативных аффиксов однозначно, имена существительные обладают способностью выполнять функцию определения.

Агглютинация - способ линейного соединения морфем в слове, состоящий в их свободном, не нарушающем морфемных границ, расположении в определённом порядке. Агглютинация в этом смысле противопоставляется фузии [4].

Сингармонизм — это морфонологическое явление, которое заключается в единообразном вокалическом (иногда и консонантном) оформлении слова как морфологической единицы [4].

Направления, в которых проявляются структурно-функциональные особенности ТЯ – это разработка приложений, управляемых знаниями. Основой таких приложений являются графы знаний (ГЗ) [5-7], которые используются для представления онтологических и фактографических знаний о мире. Это такие приложения, как приложения для определения тональности и снятия омонимии слов в задачах информационного поиска.

1 Типы графов знаний

Термин ГЗ активно используется в приложениях, управляемых знаниями. Несмотря на их распространение для решения задач разного вида, единого общепринятого определения ГЗ не существует [8]. Наиболее точным можно считать определение, представленное в работе [9]: *граф знаний* – это структурированный набор данных, собранный из разнородных источников данных, совместимый с моделью данных *RDF* и имеющий онтологию (*OWL*) в качестве своей схемы.

Формально ГЗ представляет собой граф вида $G = \{E, R, T\}$. Здесь G – размеченный ориентированный мультиграф, E – набор вершин, R – набор рёбер, T – набор триплетов вида $(u, e, v) \in T$. Где $u \in E$ – начальная вершина, $v \in E$ – конечная вершина, $e \in R$ – ребро, с началом в вершине u и концом в вершине v. В семантической интерпретации u является субъектом, v – объектом, а e – отношением между субъектом и объектом.

ГЗ называются онтологиями [10], а также определяются как системы, основанные на знаниях [11]. ГЗ позволяют описывать как отдельные предметные области, так и весь мир полностью (например, Википедия).

В качестве отдельного вида ГЗ можно выделить лингвистические ГЗ, которые описывают мир и средства для его описания в виде лингвистических единиц (ЛЕ) и структур ЕЯ. Таким образом, средства для описания мира являются метаданными по отношению к миру, но являются частью этого мира и входят в единый ГЗ (см. рисунок 1).

ГЗ языка L_i представляет собой описание языка L_i и соответствует структуре этого языка. ГЗ каждого языка L_i содержит языковые единицы этого языка, отношения между ними, а также отношения между ними и единицами ГЗ описания мира. ГЗ описания мира представляет собой определённую систему семантических универсалий, независимых от языка. Один из таких лингвистических графов, который соответствует языкам индоевропейского типа, описан в [12]. Его модель (см. рисунок 2) позволяет описывать: отношения между концепта-

ми и лексемами; информацию о встречаемости слов; диахроническую информацию об изменениях, происходящих в лексиконе.

Структура ГЗ содержит: концепты (*Concept*), лексические концепты (*Lexicon Concept*), лексемы (*Lexicon Entry*). Лексические концепты образуют единую таксономическую систему с помощью отношений гипонимии и гиперонимии. Каждой лексеме соответствуют леммы (*Lemma*) и основы словоформы (*Stem*). Такая структура соответствует языкам флективного типа, в которых лемма (словарная форма) и основа словоформы, в отличие от агглютинативных языков, не всегда совпадают. В данном ГЗ отсутствует описание типовых ситуационных фреймов, поэтому модель языка позволяет описывать только словари, а не использование слов в текстах и их потенциальные возможности.

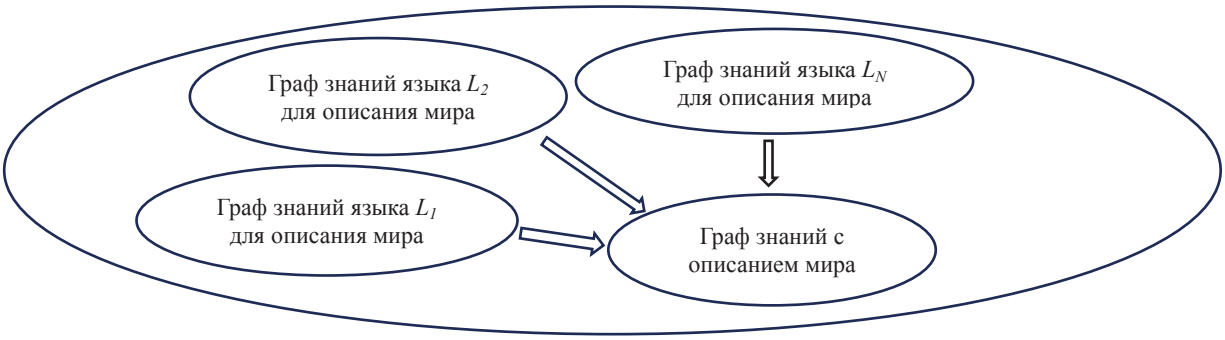

Рисунок 1 – Метаграф знаний

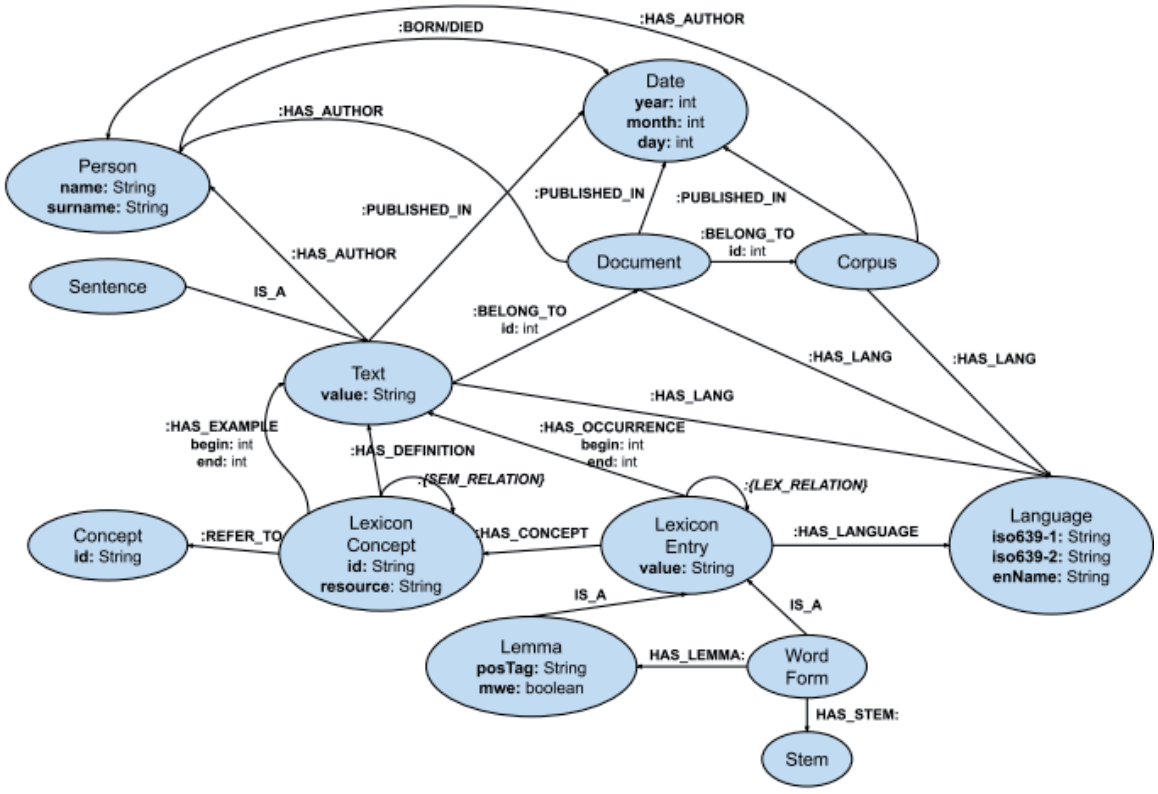

Рисунок 2 – Модель лингвистического графа знаний [11]

Возможности описания использования словоформ в текстах представлены в модели лингвистического ГЗ (*Lexicon Model for Ontologies, LeMOn* или *lemon,* см. рисунок 3) [13]. В качестве ключевого элемента в данной модели используется словарная статья (CC, *Lexical*

Entry), которая имеет наибольшее количество взаимосвязей с другими вершинами графа. В качестве СС могут быть представлены слово, фраза или часть слова. Если у какой-то лексемы есть различные варианты, например, сокращения или аббревиатура, то они представляются как отдельные СС, связанные с основной статьёй с помощью ребра графа (*lexicalVariants*). Все СС принадлежат определённым словарям (*Lexicon*), что определяется связью (*entry*). СС может состоять из нескольких лексических форм (*Lexical Form*), одна из которых помечается как каноническая форма (*canonical form*). Лексическое значение СС определяется с помощью концепта онтологии (*Lexical Sense*). Наличие в модели таких элементов, как фрейм (*Frame*) и аргумент (*Argument*) позволяют описывать семантические ситуации. СС может содержать в себе не только отдельные слова, но и многословные выражения (*Multi Word Expression, MWE*), для этого в модели представлены дополнительные компоненты (*Component*).

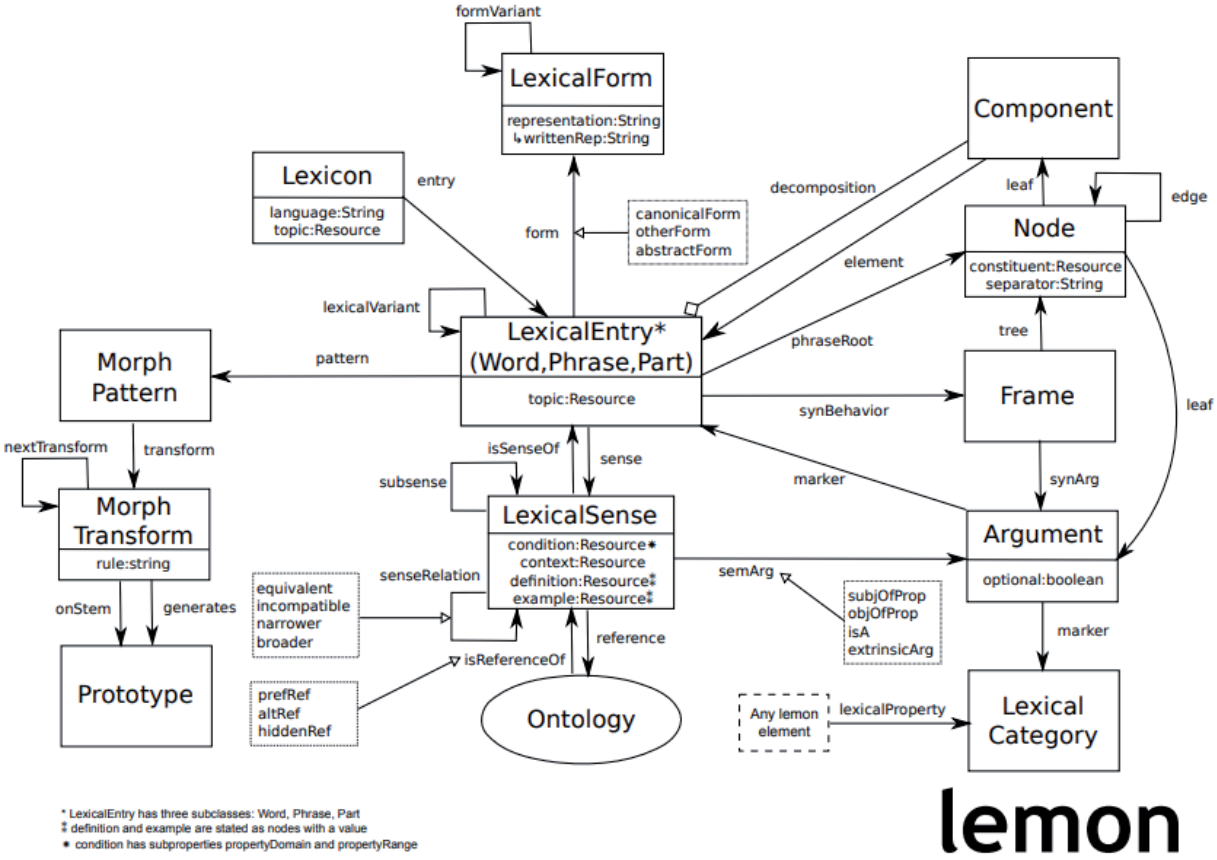

Рисунок 3 - Модель лингвистического графа знаний *Lemon* [12]

Обе модели лингвистических ГЗ позволяют описывать языки с малой морфологией типа английского. В моделях для ТЯ возникла необходимость описывать не только лексемы и словари, но также ситуационные сценарии и их выражения с помощью богатой морфологии агглютинативных языков, к которым относится семейство ТЯ.

Одной из структурных особенностей агглютинативных языков является наличие чёткого деления словоформ на морфемы. Словоформа всегда начинается с корневой морфемы, к которой справа по цепочке присоединяются аффиксальные морфемы. Эта особенность позволяет корневой морфеме выступать в роли лексемы и леммы. Аффиксальная морфема, как правило, имеет одно грамматическое значение. Данные структурные особенности позволили построить модель ГЗ ТЯ, ключевым элементом в которых является морфема.

2 Модель лингвистического ГЗ ТЯ *TurkLang*

На основе анализа существующих лингвистических графов для других типов языков, часть из которых описана в разделе 1, а также с учётом структурно-функциональных особенностей ТЯ, предложена модель ГЗ ТЯ *TurkLang*, представленная на рисунке 4.

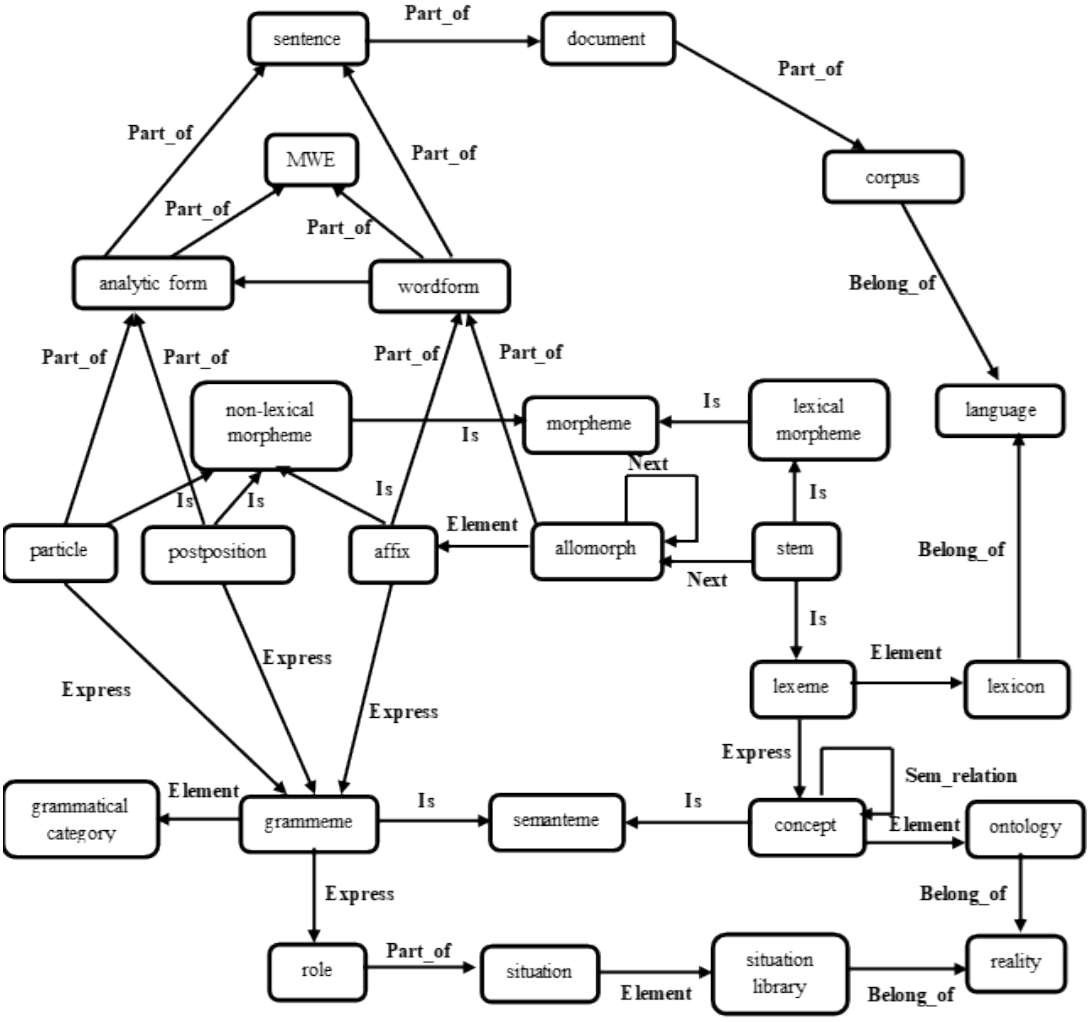

Рисунок 4 – Модель лингвистического графа знаний тюркских языков *TurkLang*[1]

Ключевым элементом модели является морфема (*morpheme*), которая может быть лексической (*lexical*) или грамматической (*non lexical*). В качестве грамматических морфем выступают аффиксальные морфемы (*affix*), послелоги (*postposition*) и частицы (*particle*). Аффиксальная морфема объединяет в себе несколько алломорфов (*allomorph*), которые представляют использование аффикса в тексте.

Лингвистическая часть графа с описанием аффиксальных морфем и более сложных лингвистических единиц, представляемая в данной модели, подробно описана в работе [3], поэтому в данной статье эти описания не приводятся. Весь перечисленный набор узлов графа образует множество ЛЕ ТЯ. Ребра между ними определяют правила образования более сложных ЛЕ. Данного набора ЛЕ и отношений достаточно для работы программ морфологического и синтаксического анализа и синтеза. Это подтверждает морфологический анализатор, реализованный в виде отдельного сервиса в рамках портала “Тюркская морфема”.

[1] На рисунке обозначения узлов графа представлены на английском языке для демонстрации аналогии с вышеописанными лингвистическим графами знаний

Для описания значений ЛЕ необходимы множества семантических универсалий, которые являются едиными для всех ТЯ. В роли таких семантических универсалий нами выбраны тезаурус, аналог *WordNet*, и модель типовых ситуационных фреймов, аналог *FrameNet*. В модели ГЗ ТЯ таксономическая часть графа представлена с помощью узлов графа концепт (*concept*) и онтология (*ontology*).

Связь языковых данных с семантическими универсалиями организована в разных лингвистических ресурсах, разрабатываемых в Институте прикладной семиотики АН РТ в лингвистическом портале «Тюркская морфема» и в электронном корпусе «Туган тел». В портале описываются все потенциально возможные варианты выражения значения концепта в разных ТЯ. Таким образом, в портале лексемы являются элементами словарей, а в электронном корпусе лексемы представляются в тексте и показывают связи при их использовании.

Концепты связаны между собой с помощью отношений гипонимии и гиперонимии (*Sem_relation*). Ситуационно-фреймовая часть семантических универсалий представлена узлами ГЗ ситуация (*situation*) и роль (*role*), которые объединяются в единую библиотеку ситуаций (*situation_library*).

Лингвистический ГЗ *TurkLang* подразделяется на несколько подграфов (см. рисунок 5). Такое разделение связано с назначением каждого из этих подграфов. Подграф семантических структур состоит из трёх основных компонентов:

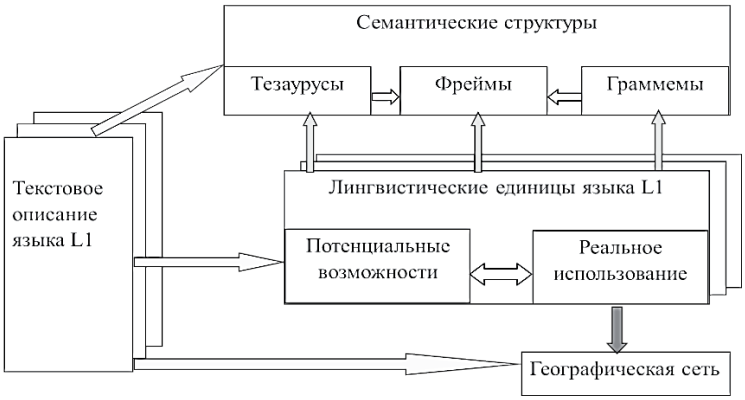

Рисунок 5 – Архитектура подграфов графа знаний *TurkLang*

- подграф тезаурусов – несколько таксономий, построенных по аналогии с известным лингвистическим ресурсом *WordNet* – подграф объектов, подграф действий, подграф атрибутов объектов и подграф атрибутов действий;
- подграф библиотеки типовых ситуационных фреймов, образующий фреймовые структуры для описания семантических типовых ситуаций и построенный по аналогии с известным лингвистическим ресурсом *FrameNet*;
- подграф граммем, узлами которой являются граммемы, связанные с ЛЕ и с библиотекой типовых ситуационных фреймов.

Лингвистические подграфы знаний соответствуют ТЯ. Тюркские языки на рисунке 5 пронумерованы: L1, L2, L3 и т.д. Их узлами являются ЛЕ данных языков разных языковых уровней: морфонологического, морфологического, синтаксического. Это корневые и аффиксальные морфемы, словоформы, аналитические формы, многословные выражения и т.д. Базовой ЛЕ в данных подграфах являются морфемы разного типа с разными наборами взаимосвязей.

Подграф географической сети содержит узлы, относящиеся к геоинформационным данным языков. На данный момент этот подграф находится в разработке в рамках новой диалектологической геолингвистической системы, соотносящей лингвистические единицы языков и текстовые описания с географическими регионами их распространения.

Представленный лингвистический граф ТЯ позволяет описывать использование ЛЕ в существующих текстах и потенциальные возможности языка, многие из которых могут находить редкое использование в текстах и речи.

Существующие тексты образуют в совокупности многоязычный электронный корпус ТЯ с возможностями создания и хранения различных типов лингвистической разметки: морфонологической, морфологической, синтаксической и разными типами семантической размет-

ки. Использование модели лингвистического ГЗ ТЯ *TurkLang* позволяет представить структуру корпуса как набор электронных корпусов каждого из ТЯ, которые взаимосвязаны между собой с помощью набора семантических универсалий. Связи в этой модели осуществляются на разных языковых уровнях с помощью библиотеки типовых ситуационных фреймов на уровне отдельных морфем и на уровне предложений,

ГЗ с описанием потенциальных возможностей ТЯ служит лингвистической БЗ для создания лингвистических процессоров для обработки текстов на разных языковых уровнях. Так, фрагменты ГЗ с описанием правил следования в словоформе используются для морфологического анализатора, типовые ситуационные фреймы - для семантико-синтаксического анализатора.

3 Программные продукты на базе модели лингвистического ГЗ ТЯ *TurkLang*

3.1 Лингвистический портал «Тюркская морфема»

Модель лингвистического ГЗ ТЯ *TurkLang* положена в основу БЗ нескольких программных продуктов, разрабатываемых в Институте прикладной семиотики Академии наук Республики Татарстан. Впервые модель *TurkLang* была использована в качестве структуры БЗ лингвистического интернет-портала «Тюркская морфема». Этот портал представляет собой сайт (*modmorph.turklang.net*), который предоставляет доступ к набору различных web-сервисов с использованием лингвистических ресурсов, структурированных на основе модели *TurkLang*. Сервисы ориентированы на компьютерную обработку ТЯ во всех аспектах: морфонологическом, морфологическом, синтаксическом, семантическом.

Сервисы можно разделить на базовые и прикладные.

- Информационно-справочная система с описанием лингвистических свойств ТЯ. Это грамматика и лексика ТЯ, представленные в виде единой модели. Лексика ТЯ представлена в виде семантического тезауруса. Справочная система структурно представляет собой некий аналог Википедии.
- ГЗ портала, как ресурсная база для лингвистических процессоров, работающих с ТЯ. Они используют разные фрагменты ГЗ.
- Набор программных модулей в виде веб-сервисов для обработки ЕЯ. Лингвистические сервисы включают лингвистические процессоры, представляющие собой анализаторы для разных языковых уровней. Основные лингвистические процессоры, реализуемые в рамках портала, – это морфологический и семантико-синтаксический анализаторы для ТЯ.
- Площадка для совместной работы и общения на тему ТЯ для специалистов, работающих с ТЯ.

Набор прикладных функций:

- инструментарий для научных исследований, например проведения сравнительных исследований ТЯ;
- инструментарий и лингвистические ресурсы для создания обучающих систем;
- инструментарий для унификации терминологии и системы тэгов для разметки электронных корпусов ТЯ.

Модель лингвистического ГЗ ТЯ *TurkLang* реализуется в БД портала в виде сущностей (см. рисунок 6):

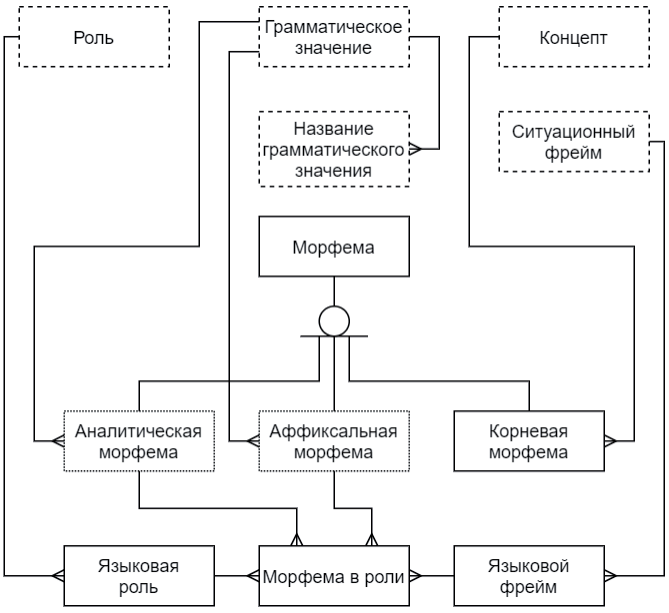

Рисунок 6 – Лингвистические подграфы знаний в базе данных портала

Аффиксальная морфема содержит информацию об аффиксах языка, выражающих грамматические значения;

Аналитическая морфема содержит информацию о частицах, послелогах и вспомогательных глаголах языка, выражающих грамматические значения;

Корневая морфема содержит информацию о корнях (леммах) языка, каждый из которых может выражать несколько значений как часть речи, связана с тезаурусом концептов;

Многословные выражения содержат информацию о неделимых словосочетаниях, выражающих целиком одно значение, связаны с тезаурусом концептов.

Языковые единицы связаны между собой морфотактическими правилами для представления морфологического уровня языка (рисунок 7).

Семантико-синтаксический уровень представлен типовыми ситуационными фреймами в языке, в которых указано, какие аффиксы требуются для выражения той или иной роли.

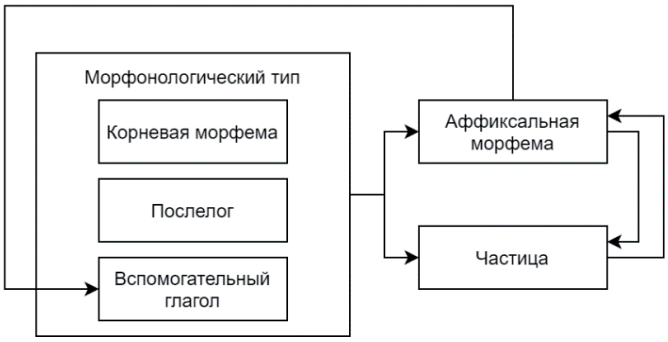

Рисунок 7 – Структура морфотактических правил

3.2 Разработка базы семантических универсалий портала «Тюркская морфема»

В модели представлено два вида семантических универсалий: тезаурус и библиотека типовых ситуационных фреймов. Анализ лингвистических ресурсов показал, что в настоящее время более полными лингвистическими ресурсами данных типов являются *WordNet* и *FrameNet*. Эти ресурсы составляют онтологическое ядро многих реальных ГЗ, вокруг которых накапливается фактографическая информация. Они приняты за основу и преобразованы с учётом структурных и лексических особенностей ТЯ.

Процесс полуавтоматического заполнения тезауруса с использованием тюркско-русских словарей, показал, что в тезаурусе *WordNet* плохо представлены разделы с описанием концептов об особенностях тюркской культуры и быта. Например, не представлены концепты, соответствующие тюркским национальным музыкальным инструментам, национальным блюдам, родственным отношениям. Данная информация была дополнена в БЗ портала. В ТЯ отсутствует ряд лексических единиц, которые соответствуют лексемам английского языка.

Тезаурус портала представлен БД концептов:

- коннекторы, выражающие союзы;
- дейктики, выражающие местоимения;
- коммуникативы, выражающие междометия и вводные слова;
- объекты выражают имена существительные, в т.ч. имена собственные;
- действия выражают глаголы;
- атрибуты объектов выражают прилагательные и причастия;
- атрибуты действий выражают наречия и деепричастия.

Концепты связаны между собой семантическими отношениями, главным из которых является иерархическое отношение, соответствующее отношению гипонимии/гиперонимии из *WordNet*.

Библиотека типовых ситуационных фреймов связывает между собой концепты (рисунок 8) по аналогии с *FrameNet* в ситуации. Каждая ситуация управляется определёнными концептами действий и состоит из набора обязательных и необязательных ролей. Данные роли могут выполняться указанными в БД концептами.

3.3 Новая версия лингвистического электронного корпуса «Туган тел»

В настоящее время разработан электронный корпус татарского языка «Туган тел», который имеет только морфологическую разметку. Существующая структура не позволяет расширить функционал корпуса для работы с синтаксической и семантической информацией.

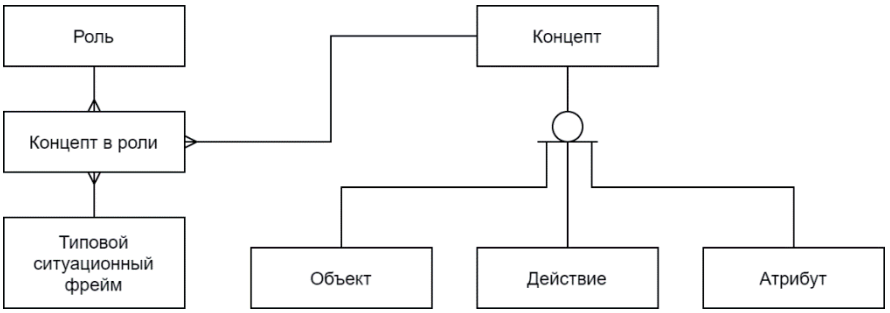

Рисунок 8 – База данных типовых ситуационных фреймов

В связи с этим начата реализация новой версии корпуса на базе модели лингвистического ГЗ ТЯ *TurkLang* с использованием графовой системой управления БД *Memgraph* (*https://memgraph.com/).* Учитывая, что лингвистический портал «Тюркская морфема» и электронный корпус «Туган тел» разрабатываются на основе общей модели, обеспечиваются совместимость этих программных продуктов, использование единой системы обозначений и тэгов. Например, в электронном корпусе используется информационно-справочная система по ТЯ портала, а портал получает из корпуса данные об использовании ЛЕ в тексте. Таким же образом интегрируются базы семантических универсалий.

На рисунке 9 представлена *ER*-диаграмма взаимодействия БД лингвистического портала «Тюркская морфема» с БД корпуса. Схема разделена на три части – общеязыковую, языкозависимую и корпусную. Общеязыковая и языкозависимая части содержат все элементы модели тюркской морфемы. Корпусная часть – это БД нового корпуса.

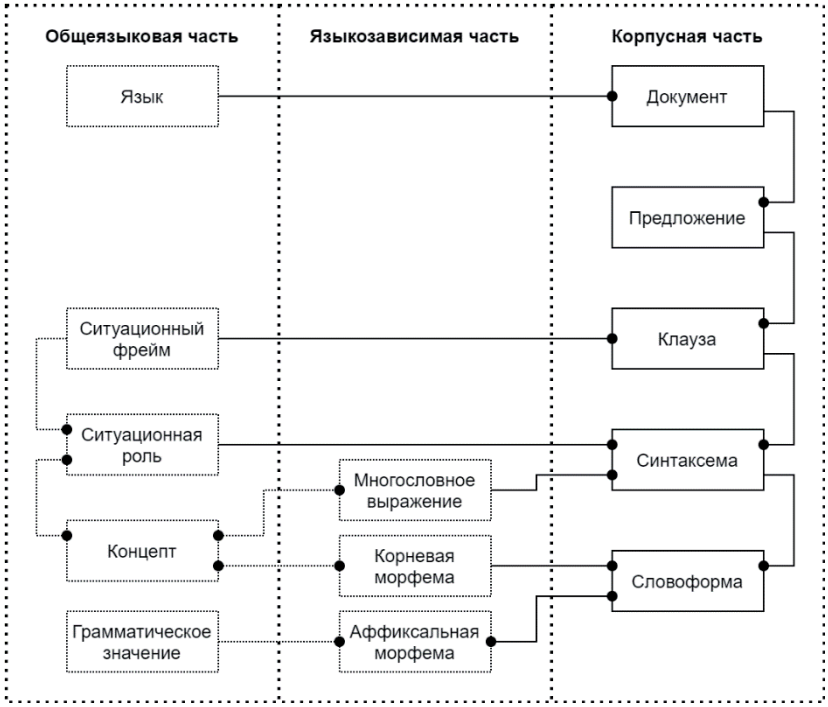

Рисунок 9 – *ER*-диаграмма объединения базы данных портала с корпусными данными

Основные сущности в корпусной части БД:

- Документ – сведения об отдельном текстовом документе корпуса;
- Предложение – данные об отдельных предложениях из текста «Документ» (могут быть простые и сложные);
- Клауза – простое предложение, состоящее из предиката (обычно выраженного глаголом) и связанных с ним аргументов;
- Синтаксема – элементарная синтаксическая единица, в рамках портала данной сущности соответствует сущность Ситуационная роль;

- Словоформа – отдельное слово, разобранное при помощи морфологического анализатора (результат анализа связывается с сущностями Корневая морфема, Аффиксальная морфема, Аналитическая морфема).

Оба программных продукта объединяются в распределённую лингвистическую платформу.

Заключение

На основе разработанной модели лингвистического ГЗ *TurkLang*, соответствующей структурно-функциональным особенностям ТЯ, создан лингвистический портал «Тюркская морфема» и осуществляется разработка новой версии электронного корпуса татарского языка «Туган тел», а также электронного диалектологического атласа ТЯ.

Данная модель является универсальной для ТЯ. Структурирование данных в ней позволяет автоматически использовать программный инструментарий портала «Тюркская морфема» или электронного корпуса «Туган тел». БД портала позволяют с помощью алгоритмов, основанных на правилах, производить аугментацию наборов данных для ТЯ, которые в дальнейшем могут использоваться для обработки ЕЯ и улучшения качества прикладных программ.

Список источников

[1] ***Krauwer S.*** The basic language resource kit (BLARK) as the first milestone for the language resources roadmap // In: Proc. International Workshop “Speech and Computer” SPECOM 2003 (Moscow, Russia, October 27-29, 2003). Moscow, 2003. P.8-15.

[2] ***Гузев В.Г.*** О некоторых экзотических особенностях тюркских языков («тюркские чудеса») // Актуальные проблемы мировой политики. 2020. Вып. 10. С.231-245. DOI: 10.21638/11701/26868318.16.

[3] ***Сулейманов Д.Ш., Гильмуллин Р.А., Гатиатуллин А.Р., Прокопьев Н.А***. Когнитивный потенциал естественных языков агглютинативного типа в интеллектуальных технологиях // Онтология проектирования. 2023. Т.13, №4(50). С.496-506. DOI:10.18287/2223-9537-2023-13-4-496-506

[4] Большая российская энциклопедия, 3-е изд., т. 1. Под ред. А.М. Прохорова. М.: Сов. энциклопедия, 1969. С.177.

[5] ***Hogan A., Blomqvist E., Cochez M., d’Amato C., de Melo G., Gutierrez C., Gayo J.E.L., Kirrane S., Neumaier S., Pollere A.*** Knowledge graphs // ACM Computing Surveys (CSUR). 2021. Vol. 54(4). P.1-37. DOI: 10.1145/3447772.

[6] ***Fensel D., Şimşek U., Angele K., Huaman E., Kärle E., Panasiuk O., Toma I., Umbrich J., Wahler A.*** Knowledge Graphs: Methodology, Tools and Selected Use Cases. Cham: Springer Cham, 2020. 164 p. DOI: 10.1007/978-3-030-37439-6.

[7] ***Ji S., Pan S., Cambria E., Marttinen P., Yu P.S.*** A Survey on Knowledge Graphs: Representation, Acquisition, and Applications // IEEE Transactions on Neural Networks and Learning Systems. 2021. Vol. 33(2). P.494-514. DOI: 10.1109/TNNLS.2021.3070843.

[8] ***Ehrlinger L., Wöß W.*** Towards a definition of knowledge graphs // In: Proc. Posters and Demos Track of 12th International Conference on Semantic Systems SEMANTiCS2016 and 1st International Workshop on Semantic Change & Evolving Semantics SuCCESS'16 co-loc. 12th International Conference on Semantic Systems SEMANTiCS 2016 (Leipzig, Germany, September 12-15, 2016). CEUR Workshop Proceedings, 2016..Vol 1695,

[9] ***Pan J.Z., Vetere G., Gomez-Perez J.M., Wu H.*** Exploiting Linked Data and Knowledge Graphs in Large Organizations. Cham: Springer Cham, 2017. 266 p. DOI: 10.1007/978-3-319-45654-6.

[10] ***Lawrynowicz A.*** Semantic data mining: an ontology-based approach. Studies on Semantic Web, vol. 29. Amsterdam: IOS Press, 2017. 194 p. DOI: 10.3233/978-1-61499-746-7-i.

[11] ***Ahmed A., Al-Masri N., Abu Sultan Y.S., Akkila A.N., Almasri A., Mahmoud A.Y., Zaqout I.S., Abu Naser S.S.*** Knowledge-based systems survey // International Journal of Academic Engineering Research (IJAER). 2019. Vol. 3(7). P.1-22.

[12] ***Basile P., Cassotti P., Ferilli S., McGillivray B.*** New Time-sensitive Model of Linguistic Knowledge for Graph Databases // In: Proc. 1st Workshop on Artificial Intelligence for Cultural Heritage co-loc. 21st International

Conference of the Italian Association for Artificial Intelligence AIxIA 2022 (Udine, Italy, November 28, 2022). CEUR Workshop Proceedings, vol. 3286, 2022. P.69-80.

[13] ***McCrae J.P., Spohr D., Cimiano P.*** Linking lexical resources and ontologies on the semantic web with Lemon // In: G. Antoniou, M. Grobelnik, E.P.B. Simperl, B. Parsia, D. Plexousakis, P.D. Leenheer, J.Z. Pan (eds.): The Semantic Web: Research and Applications. Proc. 8th Extended Semantic Web Conference ESWC 2011 Part I (Heraklion, Greece, May 29 - June 2, 2011). Lecture Notes in Computer Science, Vol. 6643. Berlin Heidelberg: Springer, 2011. P.245-259. DOI: 10.1007/978-3-642-21034-1_17.

Сведения об авторах

Гатиатуллин Айрат Рафизович, 1972 г. рождения. Окончил Казанский государственный университет в 1994 г., к.т.н. (2002). Ведущий научный сотрудник Института прикладной семиотики Академии наук Республики Татарстан. В списке научных трудов более 60 работ. ORCID: 0000-0003-3063-8147; Author ID (РИНЦ): 161758; Author ID (Scopus): 56500678000. *ayrat.gatiatullin@gmail.com* ✉.

Прокопьев Николай Аркадиевич, 1992 г. рождения. Окончил Институт вычислительной математики и информационных технологий Казанского федерального университета в 2015 году. Научный сотрудник Института прикладной семиотики Академии наук РТ. В списке научных трудов около 40 работ. ORCID: 0000-0003-0066-7465; Author ID (РИНЦ): 999214; Author ID (Scopus): 57190803409; Researcher ID (WoS): S-3829-2016. *nikolai.prokopyev@gmail.com.*

Сулейманов Джавдет Шевкетович, 1955 г. рождения. Окончил механико-математический факультет Казанского государственного университета в 1977 г., к.т.н. (1985), д.т.н. (2000). Научный руководитель Института прикладной семиотики Академии наук РТ, академик АН РТ, профессор. Заслуженный деятель науки РТ, член Российской ассоциации искусственного интеллекта. В списке научных трудов более 300 работ в области прикладной семиотики, компьютерной и когнитивной лингвистики, искусственного интеллекта. Author ID (РИНЦ): 9142; Author ID (Scopus): 6603474810; Researcher ID (WoS): B-4793-2014. *dvdt.slt@gmail.com.*

Поступила в редакцию 13.06.2024, после рецензирования 15.07.2024. Принята к публикации 22.07.2024.

Scientific article DOI: 10.18287/2223-9537-2024-14-3-366-378

Model of linguistic knowledge graphs of Turkic languages

© 2024, A.R. Gatiatullin ✉, N.A. Prokopyev, D.S. Suleymanov

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

Abstract

The article describes the TurkLang model of the linguistic knowledge graph for Turkic languages, which underpins software products for processing these languages. The model's core elements are morphemes, the smallest meaningful units of language. It captures the properties of morphemes across morphonological, morphological, syntactic, and semantic levels. This model is particularly well-suited to the structural and functional characteristics of Turkic languages, which are agglutinative, offering a comprehensive and pragmatically oriented description of their capabilities and textual manifestations. The model's properties are applied in semantic text processing software, including the "Turkic Morpheme" linguistic portal and the updated Tatar language electronic corpus "Tugan Tel." The unified TurkLang model ensures full compatibility across software products for Turkic languages and standardizes the concepts and terms used in linguistic research. This is important for Turkic languages, as many existing models are based on languages with different structures (like English or Russian) and fail to fully address the communicative, cognitive, and lexical-grammatical features of Turkic languages.

Keywords: *knowledge graph, Internet portal, linguistic unit, morpheme, Turkic language.*

For citation: *Gatiatullin AR, Prokopyev NA, Suleymanov DS.* Model of linguistic knowledge graphs of Turkic languages [In Russian]. *Ontology of designing*. 2024; 14(3): 366-378. DOI: 10.18287/2223-9537-2024-14-3-366-378.

Financial Support: This work was supported by the Russian Science Foundation (project 24-21-00453).

Conflict of interest: The authors declare no conflict of interest.

List of figures

References

[1] ***Krauwer S.*** The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In: Proc. International Workshop "Speech and Computer" SPECOM 2003 (Moscow, Russia, October 27-29, 2003). Moscow, 2003: 8-15.

[2] ***Guzev VG.*** About some exotic features of Turkic languages («Turkic miracles») [In Russian]. Digest of World Politics. 2020; 10: 231-245. DOI: 10.21638/11701/26868318.16.

[3] ***Suleymanov DS, Gilmullin RA, Gatiatullin AR, Prokopyev NA***. Cognitive potential of agglutinative languages in intelligent technologies [In Russian]. *Ontology of designing*. 2023; 13(4): 496-506. DOI:10.18287/2223-9537-2023-13-4-496-506.

[4] Great Russian Encyclopedia, 3rd ed., vol. 1. A.M. Prokhorov (ed.). Moscow.: Soviet encyclopedia, 1969. P.177, col.505.

[5] ***Hogan A, Blomqvist E, Cochez M, d'Amato C, de Melo G, Gutierrez C, Gayo JEL, Kirrane S, Neumaier S, Pollere A.*** Knowledge graphs. ACM Computing Surveys (CSUR). 2021; 54(4): 1-37. DOI: 10.1145/3447772.

[6] ***Fensel D, Şimşek U, Angele K, Huaman E, Kärle E, Panasiuk O, Toma I, Umbrich J, Wahler A.*** Knowledge Graphs: Methodology, Tools and Selected Use Cases. Cham: Springer Cham, 2020. 164 p. DOI: 10.1007/978-3-030-37439-6.

[7] ***Ji S, Pan S, Cambria E, Marttinen P, Yu PS.*** A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. IEEE Transactions on Neural Networks and Learning Systems. 2021; 33(2): 494-514. DOI: 10.1109/TNNLS.2021.3070843.

[8] ***Ehrlinger L,*** **Wöß W.** Towards a definition of knowledge graphs. In: Proc. Posters and Demos Track of 12th International Conference on Semantic Systems SEMANTiCS2016 and 1st International Workshop on Semantic Change & Evolving Semantics SuCCESS'16 co-loc. 12th International Conference on Semantic Systems SEMANTiCS 2016 (Leipzig, Germany, September 12-15, 2016). CEUR Workshop Proceedings, vol 1695, 2016.

[9] ***Pan JZ, Vetere G, Gomez-Perez JM, Wu H.*** Exploiting Linked Data and Knowledge Graphs in Large Organizations. Cham: Springer Cham, 2017. 266 p. DOI: 10.1007/978-3-319-45654-6.

[10] ***Lawrynowicz A.*** Semantic data mining: an ontology-based approach. Studies on Semantic Web, vol. 29. Amsterdam: IOS Press, 2017. 194 p. DOI: 10.3233/978-1-61499-746-7-i.

[11] ***Ahmed A, Al-Masri N, Abu Sultan YS, Akkila AN, Almasri A, Mahmoud AY, Zaqout IS, Abu Naser SS.*** Knowledge-based systems survey. *International Journal of Academic Engineering Research* (IJAER). 2019; 3(7): 1-22.

[12] ***Basile P, Cassotti P, Ferilli S, McGillivray B.*** New Time-sensitive Model of Linguistic Knowledge for Graph Databases. In: Proc. 1st Workshop on Artificial Intelligence for Cultural Heritage co-loc. 21st International Conference of the Italian Association for Artificial Intelligence AIxIA 2022 (Udine, Italy, November 28, 2022). CEUR Workshop Proceedings, vol. 3286, 2022: 69-80.

[13] ***McCrae JP, Spohr D, Cimiano P.*** Linking lexical resources and ontologies on the semantic web with Lemon // In: G. Antoniou, M. Grobelnik, E.P.B. Simperl, B. Parsia, D. Plexousakis, P.D. Leenheer, J.Z. Pan (eds.): The Semantic Web: Research and Applications. Proc. 8th Extended Semantic Web Conference ESWC 2011 Part I (Heraklion, Greece, May 29 - June 2, 2011). Lecture Notes in Computer Science, vol 6643. Berlin Heidelberg: Springer, 2011: 245-259. DOI: 10.1007/978-3-642-21034-1_17.

About the authors

Ayrat Rafizovich Gatiatullin (b. 1972) graduated from Kazan State University in 1994, PhD (2002). Leading researcher at the Institute of Applied Semiotics of Tatarstan Academy of Sciences. List of scientific works includes more than 60 works. ORCID: 0000-0003-3063-8147; Author ID (RSCI): 161758; Author ID (Scopus): 56500678000. *ayrat.gatiatullin@gmail.com* ✉.

Nikolai Arkadievich Prokopyev (b. 1992) graduated from the Institute of Computational Mathematics and Information Technologies of Kazan Federal University in 2015. Researcher at the Institute of Applied Semiotics of Tatarstan Academy of Sciences. List of scientific works includes about 40 works. ORCID: 0000-0003-0066-7465; Author ID (RSCI): 999214; Author ID (Scopus): 57190803409; Researcher ID (WoS): S-3829-2016. *nikolai.prokopyev@gmail.com.*

Dzhavdet Shevketovich Suleymanov (b. 1955) graduated from the Faculty of Mechanics and Mathematics of Kazan State University in 1977, PhD (1985), Doctor of Technical Sciences (2000). He is the scientific director at the Institute of Applied Semiotics of Tatarstan Academy of Sciences, an academician of Tatarstan Academy of Sciences, a professor, an Honored Scientist of the Republic of Tatarstan, and a member of the Russian Association of Artificial Intelligence (RAAI). The list of scientific works includes more than 300 works in the field of applied semiotics, computer and cognitive linguistics, artificial intelligence, electronic and social pedagogy. Author ID (RSCI): 9142; Author ID (Scopus): 6603474810; Researcher ID (WoS): B-4793-2014. *dvdt.slt@gmail.com.*

Received June 13, 2024. Revised July 15, 2024. Accepted July 22, 2024.