УДК 004.822 *Научная статья* DOI: 10.18287/2223-9537-2024-14-4-542-554

Электронный корпус татарского языка на базе модели лингвистических графов знаний

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

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

Аннотация

В статье представлена новая версия электронного корпуса татарского языка, модернизированная на основе модели лингвистического графа знаний тюркских языков. Новая версия корпуса позволяет описать информацию на разных лингвистических уровнях: морфонологическом, синтаксическом и семантическом благодаря представлению лингвистической информации в виде графов знаний. Такой способ представления повышает функциональные возможности работы с корпусом, позволяет производить поиск по запросам, содержащим синтаксическую и семантическую информацию. Особенность реализации электронного корпуса заключается в том, что использованная модель в наибольшей степени соответствует структурно-функциональным особенностям тюркских языков и используется в качестве основы для создания ряда программных продуктов, связанных с семантической обработкой текста на тюркских языках. В частности, к таким продуктам относятся лингвистический портал «Тюркская морфема» и новая версия электронного корпуса татарского языка «Туган тел».

Ключевые слова: *электронный корпус, граф знаний, система управления базами данных, лингвистическая единица, тюркские языки.*

Цитирование: *Гатиатуллин А.Р., Мухамедшин Д.Р., Прокопьев Н.А., Сулейманов Д.Ш.* Электронный корпус татарского языка на базе модели лингвистических графов знаний. *Онтология проектирования*. 2024. Т.14 №4(54). С. 542-554. DOI: 10.18287/2223-9537-2024-14-4-542-554.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

Современные технологии искусственного интеллекта, основанные на использовании больших языковых моделей, испытывают потребность в увеличении их информационных ресурсов за счёт включения различных электронных корпусов (ЭК). Это стало фактором усиления активности разработок ЭК для тюркских языков (ТЯ) [1-4]. В таблице 1 приведён список ЭК, проанализированных в ходе модернизации ЭК татарского языка «Туган тел». ЭК двух ТЯ включены в состав лингвистической платформы Национальный корпус русского языка (https://ruscorpora.ru/): башкирский национальный корпус объёмом 550 тыс. словоупотреблений и хакасский ЭК объёмом 1194 тыс. словоупотреблений. Большой набор ЭК собран на лингвистической платформе *Sketch Engine*, в числе которых есть тюркские ЭК (см. таблицу 2). Наибольшее количество ЭК разработано для турецкого языка, которые имеют синтаксическую или семантическую разметки: https://github.com/google-research-datasets/turkish-treebanks/ (турецкий *TreeBank*) и https://turkishpropbank.github.io/ (турецкий *PropBank*). Для турецкого языка создан лингвистический ресурс *WordNet*, с помощью которого можно организовать семантический поиск. Ресурсы для турецкого языка имеют только один вид разметки - синтаксический или семантический. В турецком *PropBank* реализована ситуационная разметка, а в *WordNet* – таксономическая. Для остальных ТЯ корпусы включают только морфологическую разметку.

Таблица 1 – Электронные корпусы тюркских языков

Название	Адрес
Башкирский поэтический корпус	http://web-corpora.net/bashcorpus/search/
Корпус башкирского языка. Проза	http://212.193.132.98/bashkorp/bashkorp
Устный корпус башкирского языка	http://lingconlab.ru/spoken_bashkir/
Алматинский корпус казахского языка	http://web-corpora.net/KazakhCorpus/search/
Национальный корпус казахского языка	http://194.146.43.249/indexru/
Национальный корпус казахского языка	https://qazcorpus.kz/about/1/?lang=ru
Крымскотатарский электронный корпус	http://korpus.juls.savba.sk/QIRIM/
Электронный корпус тувинского языка	https://www.tuvancorpus.ru/
Национальный корпус турецкого языка	https://www.tnc.org.tr/
Корпус турецкого языка	https://tscorpus.com/
Spoken Turkish Corpus	https://std.metu.edu.tr/en/
Корпус узбекского языка	https://uzbekcorpus.uz/
Электронный корпус хакасского языка	https://khakas.altaica.ru/
Корпус шорского языка	https://corpora.iea.ras.ru/corpora/
Корпус якутского языка	http://adictsakha.nsu.ru/corpora/corp
Татарский национальный корпус «Туган тел»	https://tugantel.tatar/
Письменный корпус татарского языка.	https://www.corpus.tatar/
Корпус татарской художественной литературы	http://litcorpus.antat.ru/

Таблица 2 – Электронные корпусы тюркских языков на платформе *Sketch Engine* (https://www.sketchengine.eu/)

Название	Адрес
Uzbek corpus from the web	https://www.sketchengine.eu/uzwac-uzbek-corpus/
Kazakh text corpora	https://www.sketchengine.eu/corpora-and-languages/kazakh-text-corpora/
Tatar Mixed Corpus from the web	https://www.sketchengine.eu/tatar-corpus-from-the-web/
Azerbaijani text corpora	https://www.sketchengine.eu/corpora-and-languages/azerbaijani-text-corpora/
Kyrgyz text corpora	https://www.sketchengine.eu/corpora-and-languages/kyrgyz-text-corpora/

В ЭК, размещаемых на платформе Национального корпуса русского языка, реализована возможность просмотра справочной грамматической информации о языковых единицах. Например, предыдущая версия ЭК «Туган тел» [5] включала только морфологическую разметку.

Проведённый анализ показал, что многие разработчики ЭК для ТЯ используют программный инструментарий и модели, реализованные для индоевропейского семейства языков, которые отличаются по своей структуре от ТЯ, обладающих богатой морфологией [6], а информация, представляемая в таких корпусах, не отображает всё богатство и полноту структурно-функциональных особенностей ТЯ.

Наиболее полное описание знаний и эффективное управление ими с использованием релевантных алгоритмов обработки с учётом специфики языка является важной и актуальной задачей при разработке лингвистических баз данных. Практика использования в портале «Тюркская морфема» представления данных в виде графа знаний (ГЗ) [7-9] способствует решению указанных задач, позволяя описывать в корпусе языка как онтологические, так и фактографические знания о мире.

Под ГЗ подразумевается разновидность семантической сети, определяемая в работе [10] как структурированный набор данных, собранный из разнородных источников, совместимый с моделью данных *RDF* и имеющий *OWL*-онтологию в качестве своей структуры.

Разновидностью ГЗ для представления лингвистической информации являются лингвистические ГЗ. Их отличительное свойство в том, что они описывают наряду с картиной мира также и средства для описания этого мира в виде лингвистических единиц и структур естественных языков. Исследованные в [6] лексические и грамматические особенности ТЯ [5] позволили построить модель ГЗ ТЯ, названную *TurkLang* [11]. Данная модель использовалась при создании новой версии ЭК «Туган тел».

1 Реализация архитектуры модели лингвистического ГЗ ТЯ *TurkLang* в ЭК

В проекте создания лингвистического портала «Тюркская морфема» [12] предложена модель лингвистического ГЗ ТЯ *TurkLang*, которая подходит для описания потенциальных возможностей языка и фактических данных, представленных в ЭК с текстами на ТЯ. Минимальной лингвистической единицей, представленной в этой модели, являются морфемы: корневая, аффиксальная и аналитическая. Это позволяет текст каждого предложения в корпусе представлять в виде последовательности морфем. Представление словоформы в виде фрагмента ГЗ согласно данной модели показано на рисунке 1. В узлах представлена информация о типе узла, а в скобках - содержимое конкретного узла. Узлы и рёбра фрагмента ГЗ можно условно отнести к трём уровням S1, S2, S3.

Уровень S1 – поверхностный уровень, который содержит узлы графа с информацией из реальной словоформы, использованной в тексте татарского языка.

Уровень S2 – морфемный уровень, содержит узлы ГЗ с информацией об аффиксальных морфемах татарского языка. Информация уровня S2 едина для отдельного ТЯ и узлы уровня S1 ссылаются на узлы из уровня S2.

Уровень S3 – категориальный уровень, в котором представлены узлы ГЗ, общие для всех ТЯ. Это обозначения граммем, тэгов и грамматических категорий.

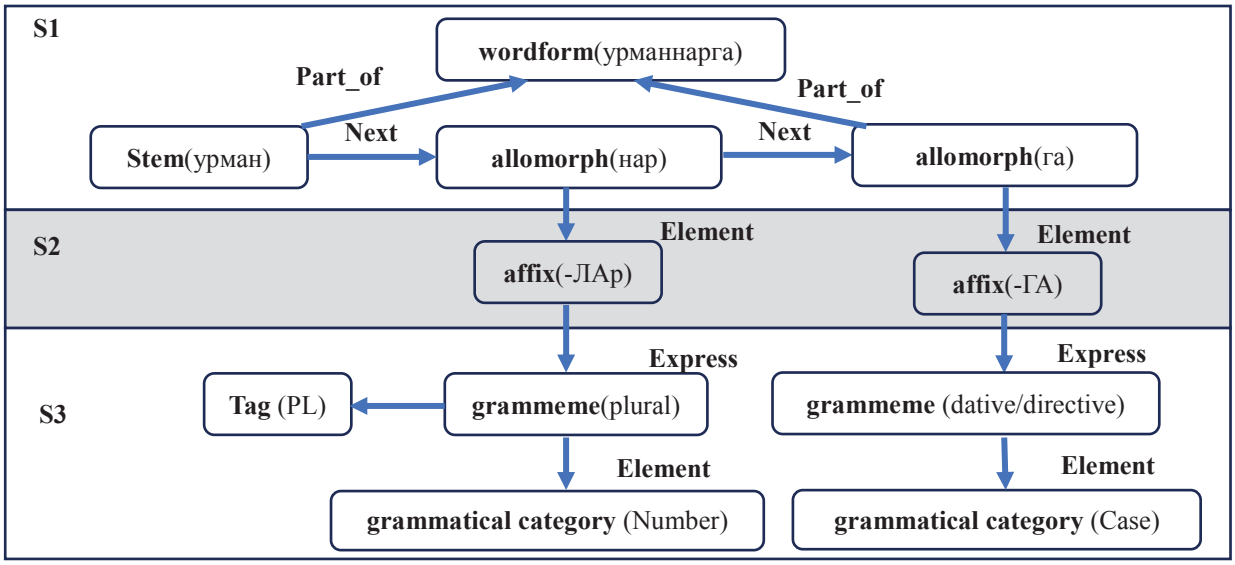


Рисунок 1 – Фрагмент графа знаний представления словоформы

Информация, представленная на уровнях S2 и S3 фрагмента ГЗ извлекается из базы знаний (БЗ) портала «Тюркская морфема», в котором специалистами по ТЯ описаны потенциальные возможности и свойства ТЯ. Такой подход позволяет использовать единую систему обозначений и обеспечить полную совместимость лингвистических ресурсов портала «Тюркская морфема» и ЭК «Туган тел». Фрагменты ГЗ с библиотеками грамматических категорий идентичны в портале и в корпусе, поэтому можно извлекать эту информацию из БЗ портала «Тюркская морфема». С целью увеличения скорости обработки поисковых запросов фрагменты ГЗ дублируются и для поддержания актуальной информации в обоих программных продуктах периодически синхронизируются.

На рисунке 1 представлен фрагмент ГЗ ЭК «Туган тел», описывающий структуру осуществления поиска в корпусе по грамматическим категориям, представленным на рисунке 2. На этом рисунке представлены все граммемы татарского языка, которые сгруппированы в грамматические категории и образуют уровень S3 ГЗ, представленного на рисунке 1.

Ещё один тип лингвистических единиц, который представлен в ГЗ ЭК «Туган тел» – это аналитические формы (*analytic form*). Аналитические формы – это формы слова с самостоя-

тельным значением в сочетании со служебными словами. Пример фрагмента ГЗ, описывающего структуру аналитической формы, представлен на рисунке 3. Аналитическими морфемами в ТЯ являются такие части речи, как послелоги, частицы или вспомогательные глаголы. Аналитические морфемы в тексте так же, как и аффиксальные морфемы, выражают грамматическую роль, что в графе определяется связью типа *Express* с узлами типа граммемы.

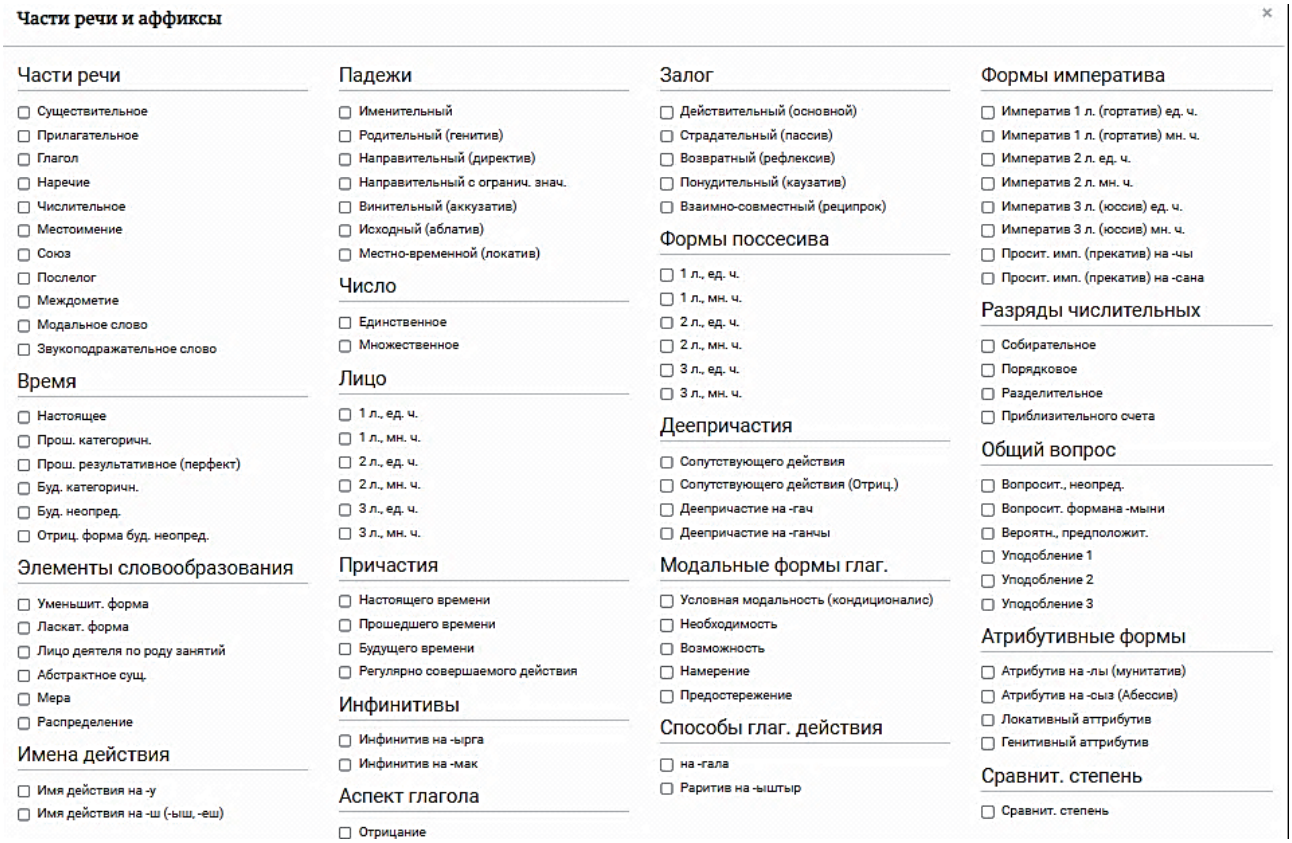


Рисунок 2 – Интерфейс для поиска в корпусе «Туган тел» по грамматическим категориям

В разных ТЯ одни и те же морфемы, выражающие одно и то же значение, могут являться как аффиксальными, так и аналитическими морфемами. Например, в татарском языке роль инструмента в тексте выражается с помощью аналитической морфемы *белән ‘с’ - чүкеч белән* ‘с молотком’, в казахском она выражается с помощью аффиксальных алломорфов ***-бен/-мен/-пен*** – *балғамен* ‘с молотком’, а в турецком с помощью аффиксальных алломорфов ***-la/-le*** – *çekiçle* ‘с молотком’. Данная особенность написания связана с различием в правилах грамматики разных ТЯ, что выражается различием в связях между узлами ГЗ, представляющих аффиксальные и аналитические алломорфы.

Графовая структура БЗ ЭК «Туган тел» позволяет хранить в БЗ семантическую, синтаксическую и морфологическую информацию, а также осуществлять семантические поисковые запросы. Для этого в БЗ ЭК хранятся подграфы с двумя видами семантических универсалий.

Первый вид – это подграф знаний с ситуационными фреймами, который является объединением ресурсов *FrameNet* (http://framenet.icsi.berkeley.edu) и *FrameBank* [13]. *FrameNet* разработан для английского языка и не учитывает морфологию лингвистических единиц, с помощью которых выражаются значения семантических универсалиев, но в нём содержится наиболее полная база типовых ситуаций. *FrameBank* создан для русского языка с формализацией грамматических структур, используемых для описания ролей в ситуационных фреймах с учётом морфологии. Поскольку ТЯ – это языки с богатой морфологией, в них необхо-

димо учитывать морфологическую информацию. Новая структура БЗ использует полноту базы *FrameNet* и морфологические элементы *FrameBank*.

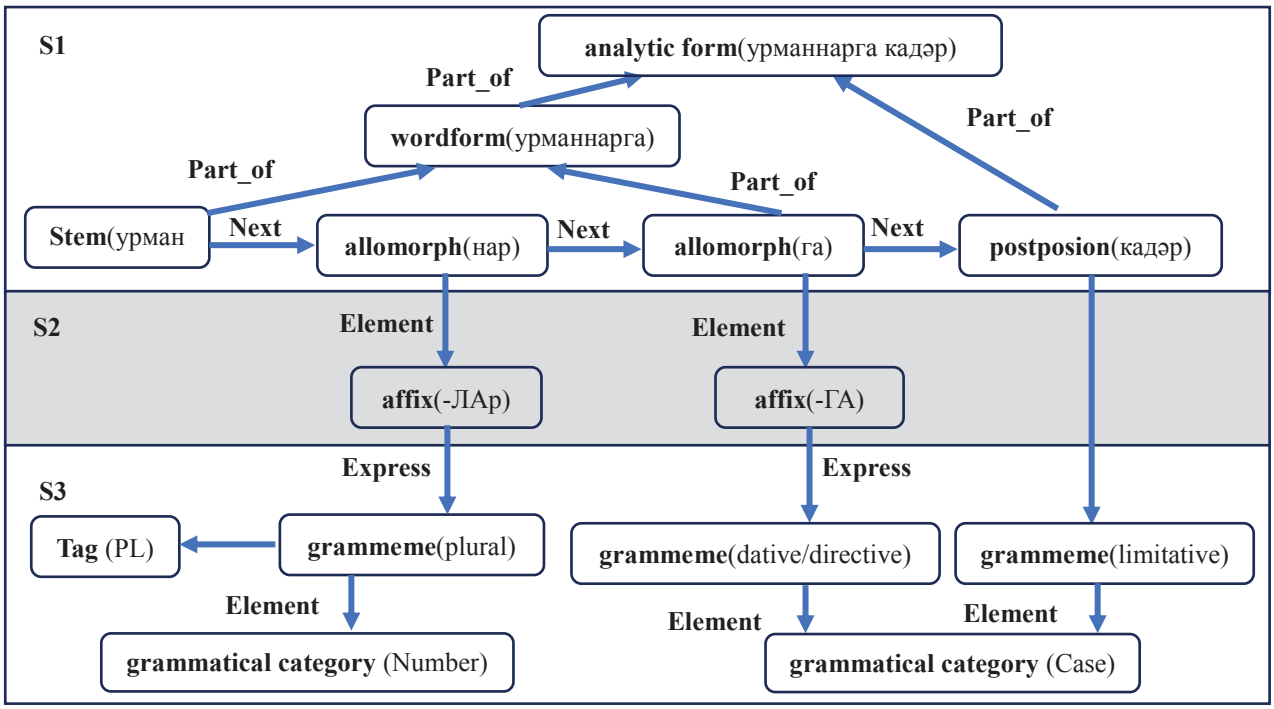


Рисунок 3 – Фрагмент графа знаний представления аналитической формы

Второй вид подграфа семантических универсалий – это таксономический подграф, реализованный в виде тезауруса типа *WordNet*. Фрагмент лингвистического ГЗ портала «Тюркская морфема» является точной копией ГЗ типа *WordNet*. Таксономическая часть графа для ТЯ представлена с помощью узлов графа концепт (*concept*), связываемых с помощью направленных рёбер. На рисунке 4 представлен фрагмент ГЗ с описанием таксономической информации, где область U ГЗ содержит семантические универсалии, которые представляют собой множество концептов и таксономические отношения между ними.

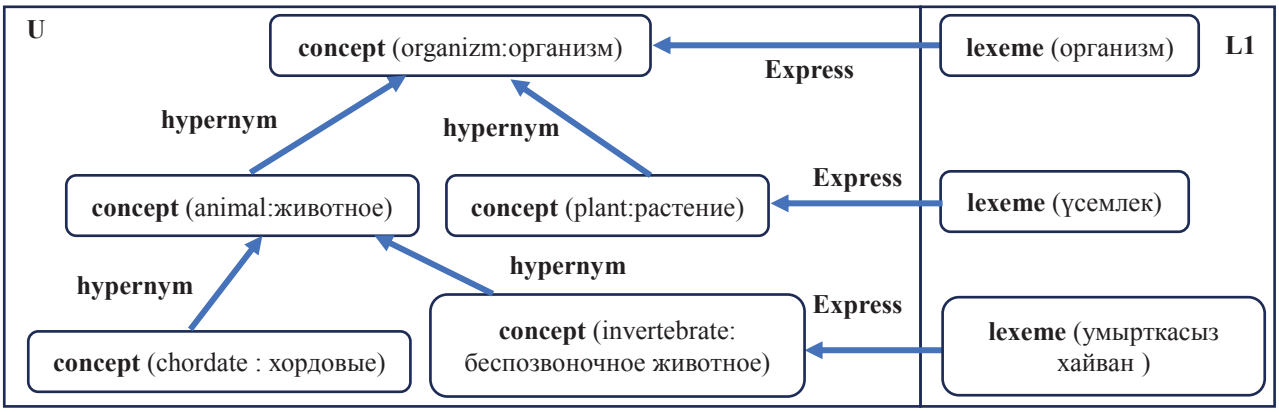


Рисунок 4 – Фрагмент графа знаний с таксономической структурой

Семантические универсалии, представленные в данной части ГЗ, в совокупности образуют семантический тезаурус. В области L1 представлены примеры лексем, которые встречаются в текстах ЭК языка (в данном примере это татарский язык). Таким образом, все лексемы *‘үсемлек’ (‘*рус.: *растение’)*, которые встречаются в корпусе, имеют связь типа

Express с концептом ‘plant:растение’. Все лексемы, которые обозначают разные виды растений, имеют связь с концептами тезауруса, которые в тезаурусе находятся с концептом ‘plant:растение’ в цепочке отношений гипонимии. Такая структура ГЗ ЭК позволяет производить семантический поиск.

Система управления корпусными данными работает с ЭК текстов на татарском языке и позволяет подключать лингвистические корпусы на других агглютинативных и флективных языках (к языкам агглютинативного типа относятся ТЯ, а к языкам флективного типа – славянские языки). Поисковые технологии реализованы на базе общедоступных программных средств: реляционной системы управления базой данных (СУБД) *MariaDB* и хранилища данных *Redis*. Для реализации предлагаемой структуры БЗ используется графовая СУБД *Memgraph*.

2 Реализация структуры БЗ ЭК «Туган тел» с помощью СУБД *Memgraph*

Первичной задачей в процессе реализации БЗ ЭК «Туган Тел» является перенос существующего ЭК в структуру ГЗ. На рисунке 5 показана итоговая схема графа, реализованная с помощью СУБД *Memgraph*, достаточная для переноса существующего ЭК в структуру БЗ.

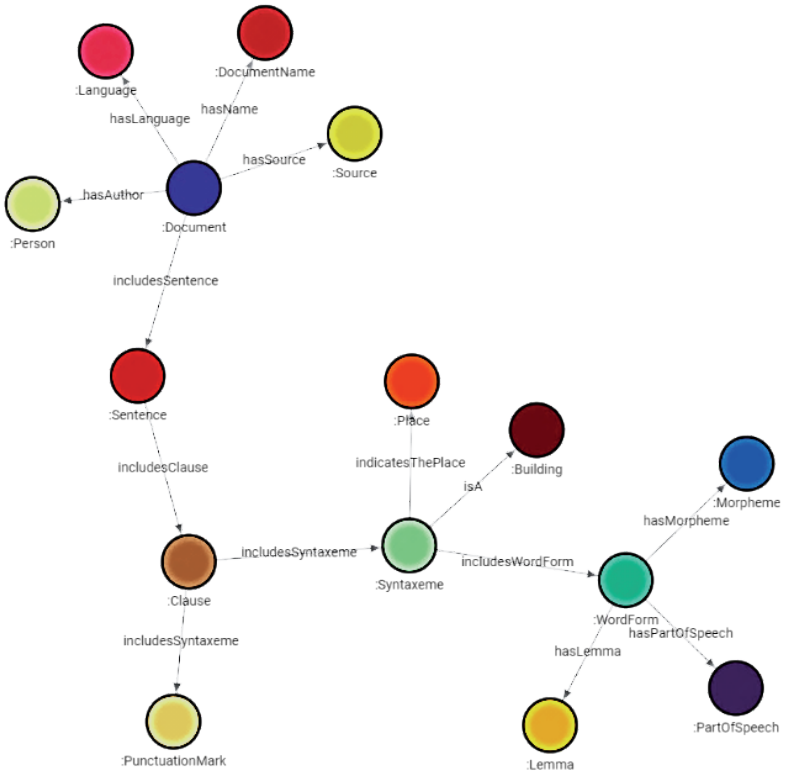


Рисунок 5 – Схема графа, реализованного с помощью СУБД *Memgraph*

В отличие от схемы, реализованной с помощью СУБД *MariaDB*, в графе дополнительно появляются узлы типов «Clause» («Клауза»), «Syntaxeme» («Синтаксема»), «PunctuationMark» («Знак препинания»), «Morpheme» («Морфема»), «PartOfSpeech» («Часть речи»), необходимых для дальнейшего представления словоформ, клауз и синтаксем. Также в графе появляются узлы «Language» («Язык»), «Person» («Человек»), «Source» («Источник»), «DocumentName» («Название документа»), «Place» («Место»), «Building» («Здание»), необходимые для дальнейшего представления семантических связей с соответствующими объектами. Количество типов таких узлов неограниченно, и их набор может быть расширен без внесения изменений в основной ГЗ.

В качестве примера в представленную структуру можно поместить предложение: «Дөрес, әле Казанда моңа кадәр картлар йорты юк иде» («И вправду, до сих пор в Казани не было дома престарелых») с морфологической разметкой, извлечённой из существующего ЭК. Для узлов типа «Sentence» предусмотрено два свойства, в которых хранятся данные о предложении в целом: «name» (предложение без морфологической разметки), «full» (предложение с морфологической разметкой). Добавление предложения осуществляется при помощи запроса на языке *Cypher*:

```
CREATE (s:Sentence {name: “Дөрес, әле Казанда моңа кадәр картлар йорты юк иде.”, full: “Дөрес (И вправду) дөрес+Adj; Type2 әле (ещё) әле+CNJ;әле+PART; Казанда (в Казани) казан+N+Sg+LOC(ДА); казан+PROP+LOC(ДА); моңа (этого) моңа+PN; кадәр (до) кадәр+Adv; кадәр+POST; картлар (старики) карт+Adj+PL(ЛАр)+Nom;карт+N+PL(ЛАр)+Nom; йорты (дом) йорт+N+Sg+POSS_3SG(СЫ)+Nom; юк юк+MOD; иде и+V+PST_DEF(ДЫ); . Type1”});
```

Далее необходимо добавить узел типа «Document». Для таких узлов предусмотрено использование трёх свойств, в которых хранятся данные о документе: «name» (наименование файла документа), «length» (длина документа в словах), «publicationDate» (дата публикации). Представление метаданных о длине документа и дате публикации в виде свойств узла обусловлено необходимостью реализации поиска по интервалам длин документов и интервалам дат. Добавление узла документа при помощи запроса на языке *Cypher* имеет вид:

```
CREATE (d:Document {name: “1_17890_1_1.txt”, length: 445, publicationDate: date(“2010-07-08”)});
```

Для добавления узлов и связей, связанных с другими метаданными документа, необходимо добавить узлы соответствующих типов («Language», «DocumentName», «Source», «Person») и рёбра между узлом документа и добавленными узлами соответствующих типов («hasLanguage», «hasName», «hasSource», «hasAuthor»). Сделать это можно одним запросом на языке *Cypher*:

```
MATCH (d:Document {name: “1_17890_1_1.txt})
MERGE (d)-[:hasLanguage]->(l:Language {name: “Tatar”})
MERGE (d)-[:hasName]->(n:DocumentName {name: “Казанда да картлар йорты ачылачак”}) («В Казани откроется дом престарелых»)
MERGE (d)-[:hasSource]->(s:Source {name: “http://www.azatliq.org/”})
MERGE (d)-[:hasAuthor]->(p:Person {name: “Наил Алан”});
```

Чтобы указать, что созданный документ включает предложение, необходимо добавить ребро между узлом документа и узлом предложения типа «includesSentence». При этом у такого ребра есть дополнительные свойства «position» (порядковый номер предложения в документе) и «startPosition» (порядковый номер первого слова предложения в документе). Так как в добавляемом примере только одно предложение, оба свойства примут значение «1». Если предложений несколько, то указанные свойства в дальнейшем помогут построить контекст вокруг предложения и найти это предложение в нужном документе. Запрос на языке *Cypher* для добавления ребра выглядит так:

```
MATCH (d:Document {name: “1_17890_1_1.txt})
MATCH (s:Sentence {name: “Дөрес, әле Казанда моңа кадәр картлар йорты юк иде.”)
MERGE (d)-[:includesSentence {position: 1, startPosition: 1}]->(s);
```

Выполнение всех описанных запросов создаёт подграф, показанный на рисунке 6. Каждое предложение в корпусе может быть разделено на клаузы. Если предложение является простым, то оно состоит из одной клаузы, сложное предложение - из двух клауз. Для добавления клауз необходимо создать узлы типа «Clause» и соединить их с узлом предложения при помощи ребра с типом «includesClause». В добавляемом предложении клауза только одна, но их может быть несколько, поэтому у рёбер типа «includesClause» должны быть указаны свойства «position» (порядковый номер клаузы в предложении) и «startPosition» (порядковый номер первого слова клаузы в предложении). Добавление клаузы при помощи запроса на языке *Cypher* может быть выполнено следующим образом:

```
MATCH (s:Sentence {name: “Дөрес, әле Казанда моңа кадәр картлар йорты юк иде.”)
MERGE (s)-[:includesClause {position: 1, startPosition: 1}]->(c:Clause {name: “Дөрес, әле Казанда моңа кадәр картлар йорты юк иде.”});
```

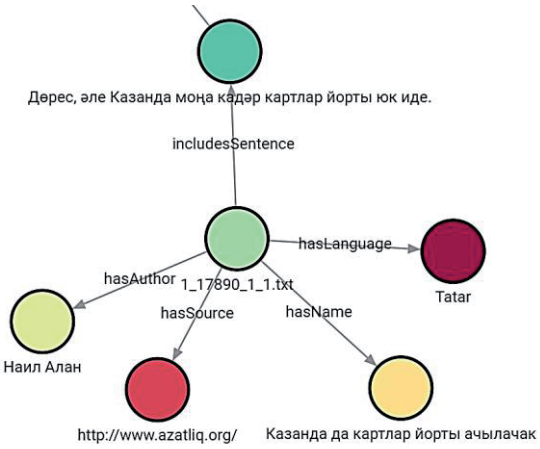


Рисунок 6 – Подграф, включающий узлы предложения, документа и метаданных документа

Каждая клауза в ЭК может быть разделена на синтаксемы. Синтаксема - это минимальная, неделимая семантико-синтаксическая языковая единица, выступающая одновременно как носитель элементарного смысла и как конструктивный компонент более сложных синтаксических построений. Синтаксеме может соответствовать как отдельная словоформа, так и словосочетание или знак препинания. Таким образом, для представления синтаксем в БЗ используются узлы типов «Syntaxeme» для синтаксем, состоящих из словоформ, и «PunctuationMark» для синтаксем, состоящих из знаков препинания. Для представления связей между клаузами и синтаксемами используются рёбра типа «includesSyntaxeme», у которых должны быть указаны

свойства «position» (порядковый номер синтаксемы в клаузе) и «startPosition» (порядковый номер первой словоформы или знака препинания синтаксемы в клаузе). Запрос для добавления синтаксем на языке *Cypher* представлен ниже:

```
MATCH (c:Clause {name: “Дөрес, әле Казанда моңа кадәр картлар йорты юк иде.”)
MERGE (c)-[:includesSyntaxeme {position: 1, startPosition: 1}]->(:Syntaxeme {name: “дөрес”})
MERGE (c)-[:includesSyntaxeme {position: 2, startPosition: 2}]->(:PunctuationMark {name: “,”})
MERGE (c)-[:includesSyntaxeme {position: 3, startPosition: 3}]->(:Syntaxeme {name: “әле”})
MERGE (c)-[:includesSyntaxeme {position: 4, startPosition: 4}]->(:Syntaxeme {name: “казанда”})
MERGE (c)-[:includesSyntaxeme {position: 5, startPosition: 5}]->(:Syntaxeme {name: “моңа кадәр”})
MERGE (c)-[:includesSyntaxeme {position: 6, startPosition: 7}]->(:Syntaxeme {name: “картлар йорты”})
MERGE (c)-[:includesSyntaxeme {position: 7, startPosition: 9}]->(:Syntaxeme {name: “юк иде”})
MERGE (c)-[:includesSyntaxeme {position: 8, startPosition: 11}]->(:PunctuationMark {name: “.”});
```

Выполнение запросов на добавление клауз и синтаксем создаёт подграф, показанный на рисунке 7. Узлы типа «PunctuationMark» являются конечными в текущей версии БЗ. Синтаксемы, состоящие из словоформ, должны быть разделены на словоформы. Словоформы представлены в графе БЗ узлами типа «WordForm», а связи между синтаксемами и словоформами – рёбрами типа «includesWordForm» со свойствами «position», указывающими порядковый номер словоформы в синтаксеме. В качестве примера показаны запросы для синтаксем “Казанда” («в Казани») и “картлар йорты” («дом престарелых»). Запрос, добавляющий в граф БЗ словоформы и связи с указанными синтаксемами, на языке *Cypher* выглядит следующим образом:

```
MATCH (s1:Syntaxeme {name: “казанда”})
MATCH (s2:Syntaxeme {name: “картлар йорты”})
MERGE (s1)-[:includesWordForm {position: 1}]->(:WordForm {name: “казанда”})
MERGE (s2)-[:includesWordForm {position: 1}]->(:WordForm {name: “картлар”})
MERGE (s2)-[:includesWordForm {position: 2}]->(:WordForm {name: “йорты”});
```

Морфологическая разметка каждой словоформы содержит лемму, часть речи и набор морфологических свойств (морфем) словоформы. Причём в части корпуса у каждой словоформы может быть несколько вариантов морфологической разметки.

Для представления в БЗ лемм используются узлы типа «Lemma», для представления частей речи – узлы типа «PartOfSpeech», для представления морфологических свойств – узлы типа «Morpheme». Последние имеют справочное свойство «affix», в котором указывается словообразующий аффикс, соответствующий морфеме. Связи между узлами словоформ и узлами лемм представлены в графе рёбрами типа «hasLemma», связи между узлами словоформ и узлами частей речи – рёбрами типа «hasPartOfSpeech», а связи между узлами словоформ и узлами морфем – рёбрами типа «hasMorpheme». Так как в ЭК может иметься разметка с морфологической неоднозначностью, у всех указанных рёбер присутствуют свойства «variant», указывающие на порядковый номер морфологической разметки словоформы. Для рёбер типа «hasMorpheme» дополнительно указывается свойство «position», указывающее на порядковый номер морфемы в цепочке. Добавление указанных узлов и рёбер в граф БЗ при помощи запроса на языке *Cypher* может быть представлено следующим образом:

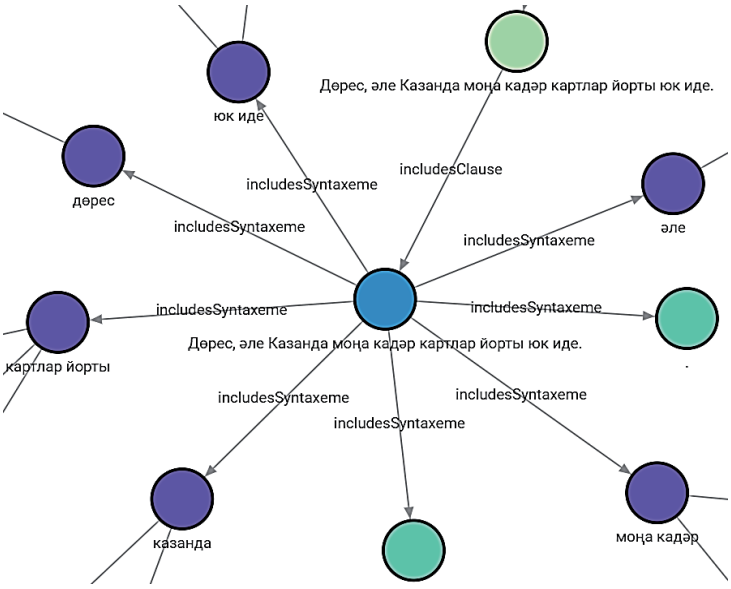


Рисунок 7 – Подграф, включающий узлы предложения, клауз и синтаксем

```
MATCH (w1:WordForm {name: “казанда”})
MATCH (w2:WordForm {name: “картлар”})
MATCH (w3:WordForm {name: “йорты”})
```

```
CREATE (:Morpheme {name: “SG”, affix: “”}), (:Morpheme {name: “LOC”, affix: “да”}), (:Morpheme {name: “PL”, affix: “ЛАр”}), (:Morpheme {name: “NOM”, affix: “”}), (:Morpheme {name: “POSS_3SG”, affix: “СЫ”})
MERGE (w1)-[:hasLemma {variant: 1}]->(:Lemma {name: “казан”})
MERGE (w1)-[:hasPartOfSpeech {variant: 1}]->(:PartOfSpeech {name: “N”})
MERGE (w1)-[:hasMorpheme {variant: 1, position: 1}]->(:Morpheme {name: “SG”})
MERGE (w1)-[:hasMorpheme {variant: 1, position: 2}]->(:Morpheme {name: “LOC”})
MERGE (w1)-[:hasLemma {variant: 2}]->(:Lemma {name: “казан”})
MERGE (w1)-[:hasPartOfSpeech {variant: 2}]->(:PartOfSpeech {name: “PROP”})
 MERGE (w1)-[:hasMorpheme {variant: 2, position: 1}]->(:Morpheme {name: “LOC”})
MERGE (w2)-[:hasLemma {variant: 1}]->(:Lemma {name: “карт”})
MERGE (w2)-[:hasPartOfSpeech {variant: 1}]->(:PartOfSpeech {name: “ADJ”})
MERGE (w2)-[:hasMorpheme {variant: 1, position: 1}]->(:Morpheme {name: “PL”)
MERGE (w2)-[:hasMorpheme {variant: 1, position: 2}]->(:Morpheme {name: “NOM”)
MERGE (w2)-[:hasLemma {variant: 2}]->(:Lemma {name: “карт”})
MERGE (w2)-[:hasPartOfSpeech {variant: 2}]->(:PartOfSpeech {name: “N”})
MERGE (w2)-[:hasMorpheme {variant: 2, position: 1}]->(:Morpheme {name: “PL”)
MERGE (w2)-[:hasMorpheme {variant: 2, position: 2}]->(:Morpheme {name: “NOM”)
MERGE (w3)-[:hasLemma {variant: 1}]->(:Lemma {name: “йорт”})
MERGE (w3)-[:hasPartOfSpeech {variant: 1}]->(:PartOfSpeech {name: “N”})
MERGE (w3)-[:hasMorpheme {variant: 1, position: 1}]->(:Morpheme {name: “SG”)
MERGE (w2)-[:hasMorpheme {variant: 1, position: 2}]->(:Morpheme {name: “POSS_3SG”);
```

Необходимо также добавить семантические связи между лексемами и объектами. Указанные синтаксемы обозначают здание “картлар йорты” (“дом престарелых”) (в графе узел типа «Building») и место “Казан” (“Казань”) (в графе узел типа “Place”). Добавление семантических связей при помощи запроса на языке *Cypher* выглядит так:

```
MATCH (s1:Syntaxeme {name: “казанда”})
MATCH (s2:Syntaxeme {name: “картлар йорты”})
MERGE (s1)-[:indicatedThePlace]->(:Place {name: “Казан”})
MERGE (s2)-[:isA]->(:Building {name: “картлар йорты”});
```

На рисунке 8 показан подграф, созданный после выполнения указанных запросов для синтаксем “Казанда” и “картлар йорты”. На этом подграфе уже видны общие вершины графа для различных типов узлов и рёбер, например, все три словоформы из примера имеют связь с частью речи “N” (существительное), так как для всех трёх словоформ существует морфологическая разметка с указанием этой части речи.

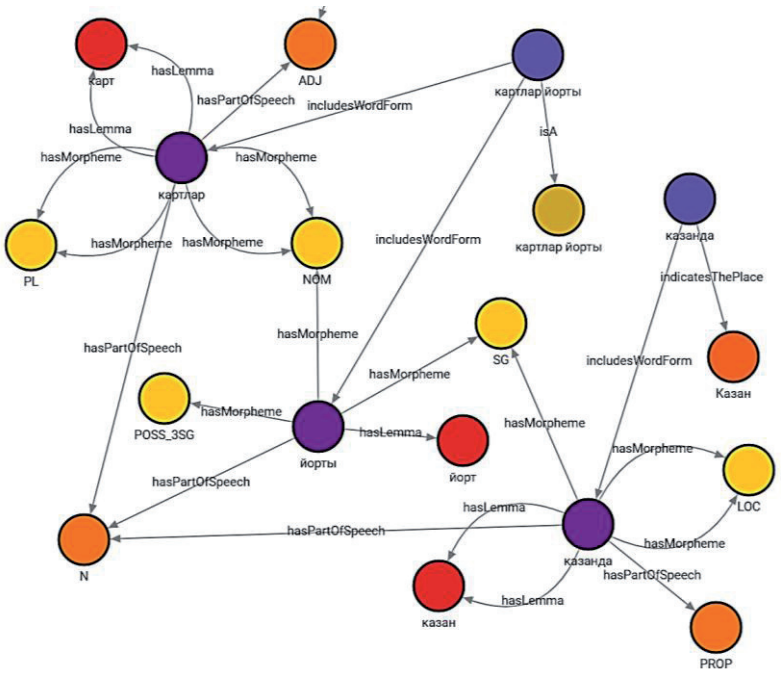


Рисунок 8 – Подграф, включающий узлы синтаксем, словоформ, лемм, морфем, частей речи и семантических связей с объектами

Заключение

Описанная концептуальная модель лингвистического ГЗ применяется для представления данных ЭК «Туган Тел». Реализация функционала модернизированной системы управления корпусными данными поддерживает функционал поиска по словоформам и леммам, а также

поиск по морфемам. Применение новой модели лингвистического ГЗ и возможностей графовой СУБД позволяет расширить функционал системы, добавляя новые инструменты для исследования ТЯ.

Использование системы семантических универсалий в виде фреймовых и таксономических ГЗ позволяет объединять все корпусы в единый многоязычный корпус и производить многоязычный поиск и исследования ТЯ. Эта возможность позволит повысить эффективность деятельности лингвистов и типологов, работающих с ЭК на основе предложенной модели лингвистического ГЗ ТЯ *TurkLang*.

Список источников

[1] ***Aksan M., Aksan Y.*** Linguistic Corpora: A View from Turkish. In: Oflazer, K., Saraçlar, M. (eds) Turkish Natural Language Processing. Theory and Applications of Natural Language Processing. 2018. Springer, Cham. DOI:10.1007/978-3-319-90165-7_14.

[2] ***Салчак А.Я.*** Электронный корпус текстов тувинского языка. *Новые исследования Тувы*. 2012. №3. С.110-114.

[3] ***Bazarbayeva Z.M., Zharkynbekova Sh.K., Amanbayeva A.Zh., Zhumabayeva Zh.T., Karshygayeva A.A.*** The National Corpus of Kazakh Language: Development of Phonetic and Prosodic Markers. *Journal of Siberian Federal University. Humanities and Social Sciences*. 2023. T. 16. № 8. P.1256-1270. EDN: IVPVAN.

[4] ***Sirazitdinov, Z. Buskunbaeva L., Ishmukhametova A.*** About linguistic corpora of the Bashkir language // Proceedings of the International Conference "Turkic languages processing" Turklang-2015 / Tatarstan Academy of Sciences L.N. Gumilyov Eurasian National University Ministry of Education and Science of the Republic of Kazakhstan Kazan Federal University Institute of Philology and Intercultural Communication. – Казань, Россия: Академия наук Республики Татарстан, 2015. P.269-275. EDN ZDGYTR.

[5] ***Mukhamedshin D., Gilmullin R., Khakimov B.*** Search Engine Capabilities in the Corpus Data Management System // UBMK 2023 - Proceedings: 8th International Conference on Computer Science and Engineering, Burdur; Turkey; 13-15 September 2023, p.449–452. DOI: 10.1109/UBMK59864.2023.10286648.

[6] ***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.

[7] ***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.

[8] ***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.

[9] ***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.

[10] ***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.

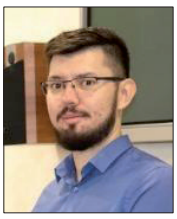
[11] ***Гатиатуллин А.Р., Прокопьев Н.А., Сулейманов Д.Ш.*** Модель лингвистических графов знаний тюркских языков // *Онтология проектирования*. 2024. Т.14, №3(53). С.366-378. DOI: 10.18287/2223-9537-2024-14-3-366-378

[12] ***Gatiatullin A., Suleymanov D., Prokopyev N., Khakimov B.*** About turkic morpheme portal // *CEUR Workshop Proceedings*, 2020, 2780. P.226–243. EDN: ZNIQUO.

[13] ***Lyashevskaya, O. and Egor Kashkin***, FrameBank: A Database of Russian Lexical Constructions // International Joint Conference on the Analysis of Images, Social Networks and Texts, 2015. M.Y. Khachay et al. (Eds): AIST 2015, CCIS 542. P.1–11. DOI: 10.1007/978-3-319-26123-2_34.

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 90 работ. ORCID: 0000-0003-3063-8147; Author ID (РИНЦ): 161758; Author ID (Scopus): 56500678000. *ayrat.gatiatullin@gmail.com*✉



Мухамедшин Дамир Рафкатович, 1993 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 более 20 работ. ORCID: 0000-0003-0078-9198; Author ID (РИНЦ): 1031142; Author ID (Scopus): 57194654368; Researcher ID (WoS): KPY-5366-2024. *damirmuh@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. (1986), 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 (РАИИ). В списке научных трудов более 300 работ в области прикладной семиотики, компьютерной и когнитивной лингвистики, искусственного интеллекта, электронной и социальной педагогики. 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 4.10.2024. Принята к публикации 28.10.2024.

CC BY

Scientific article DOI: 10.18287/2223-9537-2024-14-4-542-554

Electronic corpus of the Tatar language based on the model of linguistic knowledge graphs

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

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

Abstract

The article presents a new version of the electronic corpus of the Tatar language, updated based on a linguistic knowledge graph model for Turkic languages. This new version of the corpus allows for information description across multiple linguistic levels: morphonological, syntactic, and semantic, through the use of knowledge graphs to represent linguistic data. This approach enhances corpus functionality, enabling searches that incorporate syntactic and semantic information. A distinctive feature of the electronic corpus implementation is that the model employed aligns closely with the structural and functional characteristics of Turkic languages and serves as a foundation for developing various software products for semantic text processing in Turkic languages. In particular, these products include the linguistic portal "Turkic Morphme" and the new version of the Tatar language electronic corpus, "Tugan Tel.".

Keywords: *electronic corpus, knowledge graph, database management system, linguistic unit, turkic languages.*

For citation: *Gatiatullin AR, Mukhamedshin DR, Prokopyev NA, Suleymanov DS.* Electronic corpus of the Tatar language based on the model of linguistic knowledge graphs [In Russian]. *Ontology of designing*. 2024; 14(4): 542-554. DOI: 10.18287/2223-9537-2024-14-4-542-554.

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

List of figures and tables

References

[1] ***Aksan M, Aksan Y.*** Linguistic Corpora: A View from Turkish. In: Oflazer, K., Saraçlar, M. (eds) Turkish Natural Language Processing. Theory and Applications of Natural Language Processing. 2018. Springer, Cham. DOI:10.1007/978-3-319-90165-7_14.

[2] ***Salchak AYa.*** Electronic corpus of texts of the Tuvan language [In Russian]. *The New Research of Tuva*. 2012; 3(15): 110-114.

[3] ***Bazarbayeva ZM, Zharkynbekova ShK, Amanbayeva AZh, Zhumabayeva ZhT, Karshygayeva AA.*** The National Corpus of Kazakh Language: Development of Phonetic and Prosodic Markers // *Journal of Siberian Federal University. Humanities and Social Sciences*. 2023; 16(8): 1256-1270. EDN: IVPVAN.

[4] ***Sirazitdinov Z., Buskunbaeva L., Ishmukhametova A.*** About linguistic corpora of the Bashkir language // Proceedings of the International Conference "Turkic languages processing" Turklang-2015 / Tatarstan Academy of Sciences L.N. Gumilyov Eurasian National University Ministry of Education and Science of the Republic of Kazakhstan Kazan Federal University Institute of Philology and Intercultural Communication. – Kazan, Russia: Tatarstan Academy of Sciences, 2015. P.269-275. EDN ZDGYTR.

[5] ***Mukhamedshin D., Gilmullin R., Khakimov B.*** Search Engine Capabilities in the Corpus Data Management System // UBMK 2023 - Proceedings: 8th International Conference on Computer Science and Engineering, Burdur; Turkey; 13-15 September 2023, pp. 449–452. DOI: 10.1109/UBMK59864.2023.10286648.

[6] ***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.

[7] ***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.

[8] ***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.

[9] ***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.

[10] ***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.

[11] ***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.

[12] ***Gatiatullin A, Suleymanov D, Prokopyev N, Khakimov B.*** About turkic morpheme portal // *CEUR Workshop Proceedings*, 2020; 2780: 226–243. EDN: ZNIQUO.

[13] ***Lyashevskaya O, Kashkin E.*** FrameBank: A Database of Russian Lexical Constructions // International Joint Conference on the Analysis of Images, Social Networks and Texts, 2015. M.Y. Khachay et al. (Eds): AIST 2015, CCIS 542. P.1–11. DOI: 10.1007/978-3-319-26123-2_34.

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*✉

Damir Rafkatovich Mukhamedshin (b. 1993) 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 more than 20 works. ORCID: 0000-0003-0078-9198; Author ID (RSCI): 1031142; Author ID (Scopus): 57194654368; Researcher ID (WoS): KPY-5366-2024. *damirmuh@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). Scientific Director of the Institute of Applied Semiotics of Tatarstan Academy of Sciences, Academician of Tatarstan Academy of Sciences, Professor. Honored Scientist of the Republic of Tatarstan, member of the Russian Association of Artificial Intelligence (RAAI). List of scientific works includes more than 300 works in the field of applied semiotics, computer and cognitive linguistics, artificial intelligence, and 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 October 4, 2024. Accepted October 28, 2024.