УДК 004.822

МОДЕЛЬ ЛИНГВИСТИЧЕСКОГО ГРАФА ЗНАНИЙ «TURKLANG» КАК БАЗА ДЛЯ СОЗДАНИЯ ИНСТРУМЕНТОВ ОБУЧЕНИЯ ТЮРКСКИМ ЯЗЫКАМ

А. Р. Гатиатуллин[1] [0000-0003-3063-8147], Н. А. Прокопьев[2] [0000-0003-0066-7465]

[1]*Академия наук РТ;* [2]*Казанский Федеральный Университет*

[1]ayrat.gatiatullin@gmail.com, [2]nikolai.prokopyev@gmail.com

Аннотация

Описаны элементы модели лингвистического графа знаний «Turklang», разработанного в Институте прикладной семиотики АН РТ и используемого в качестве базы для создания ряда лингвистических ресурсов и инструментов: портал «Тюркская морфема», электронный корпус татарского языка «Туган Тел», лингвистические процессоры.

Для создания образовательной среды необходимы предметно-ориентированные графы знаний, для получения которых не применимы методы создания общих и открытых графов. В работе описаны лингвистические графы знаний, которые отображают, с одной стороны, потенциальные возможности тюркских языков, с другой стороны, примеры реального использования в текстах. Особенность этих графов знаний заключается в том, что они содержат лингвистические единицы разных языковых уровней, а также семантические универсалии, соответствующие значениям этих лингвистических единиц, которые встроены в единую модель лингвистического графа знаний. Структура такого графа знаний позволяет формировать учебные курсы, строить индивидуальную образовательную траекторию, а также создавать задания и средства автоматизированной проверки в рамках контроля знаний при обучении тюркским языкам. Это дает возможность разрабатывать впоследствии, на основе этих графов, программы обучения с учетом структурно-функциональных особенностей тюркских языков, а также способствует реализации индивидуальных целей обучающихся.

Ключевые слова*: граф знаний, база знаний, лингвистический ресурс, лингвистическая единица, малоресурсные языки, тюркские языки, веб-портал, электронное образование, контроль знаний, автоматизированная оценка ответа*

© А. Р. Гатиатуллин, Н. А. Прокопьев, 2024.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

ВВЕДЕНИЕ

В 1991 году Т. Джонс [1] выдвинул идею обучения языку на основе лингвистических баз данных и гипотезу, что обучение языкам будет более эффективным, если обучающийся сам будет выступать в роли исследователя языка, а учитель будет обеспечивать ему контекст и направления познания языка. Таким образом, учащиеся имеют возможность работать с лингвистическими базами данных и проводят исследование в своих учебных целях. Т. Джонс выделяет следующие преимущества данного подхода к обучению:

1. Обучающийся учится видеть языковые структуры, искать аналогии и обобщать полученные данные. Кроме того, использование аутентичных материалов делает речь обучающихся более идиоматичной, приближая ее к речи носителей языка.
2. Преподаватель вместо транслятора информации о языке становится координатором исследований ученика, что позволяет ученику самостоятельно обрабатывать потоки информации «об особенностях употребления языковой формы, а также решать задачи, связанные с ее осознанием».
3. Изменяется роль грамматики в изучении иностранного языка. Утверждается, что грамматика языка не способна отразить все разнообразие его синтаксических структур, поэтому неэффективно изучать грамматику отдельно от области функционирования правил грамматики. Использование в учебном процессе лингвистических ресурсов может сделать этот процесс более естественным.

Мы считаем, что приведенные положения справедливы и для тюркских языков, которые обладают богатой морфологией, и для проверки названной гипотезы необходимо наличие лингвистических ресурсов, содержащих тюркские лингвистические базы знаний. С учетом структурной близости тюркских языков эти ресурсы могут быть универсальными для всех тюркских языков. Свою положительную роль может сыграть и многоязычность этих ресурсов, потому что обучаемый сможет сравнивать языковые примеры в разных тюркских языках, акцентируя внимание на особенностях конкретного изучаемого языка.

В работе [2] авторы используют идеи, выдвинутые Т. Джонсом [1], и выделяют три вида учебных материалов, которые обучающиеся могут использовать для изучения иностранного языка:

1. Электронная лексикография;
2. Корпусные исследования;
3. Проектирование электронных учебников и учебных терминологических баз данных.

К преимуществам изучения иностранного языка с использованием перечисленных учебных материалов, с учетом идей Т. Джонса, авторы [2] добавили ряд дополнительных пунктов:

1. Работа с актуальным материалом;
2. Развитие исследовательских навыков;
3. Усвоение естественного построения речи.

Авторы работы [2] также считают, что ценность лингвистической базы данных при изучении иностранного языка повышается, если она сочетает в себе учебную, справочную, систематизирующую и коммуникативную функции.

Воспользуемся гипотезами, выдвинутыми авторами перечисленных работ, для их апробации на тюркских языках. Для проверки сформулированных утверждений необходимо наличие лингвистических ресурсов, которые, как указано ранее, должны содержать в себе учебную, справочную, систематизирующую и коммуникативную функции. Однако в настоящее время практически все тюркские языки, кроме турецкого, являются малоресурсными языками, так как испытывают недостаток в лингвистических ресурсах различных типов. В Институте прикладной семиотики Академии наук Республики Татарстан разрабатывается ряд лингвистических ресурсов для компьютерной обработки тюркских языков. В роли наиболее значимых из них можно выделить:

1. Лингвистический портал «Тюркская морфема»;
2. Электронный корпус «Туган тел».

Электронный корпус «Туган тел» ранее был реализован в двух версиях на разных технологических платформах, в настоящее время создается третья версия электронного корпуса на базе графовых баз данных. Графовые базы данных позволяют более эффективно представлять в корпусе синтаксическую и семантическую информацию.

В основе новой версии электронного корпуса «Туган тел» и лингвистического портала «Тюркская морфема» лежит единая лингвистическая модель знаний «TurkLang», реализованная в виде лингвистического графа знаний. Эта модель также является разработкой Института прикладной семиотики АН РТ.

ЛИНГВИСТИЧЕСКИЕ ГРАФЫ ЗНАНИЙ

В настоящее время одним из эффективных способов представления лингвистической информации в различных ресурсах являются графы знаний. Имеется целый ряд работ с описанием лингвистических графов знаний, в основном зарубежных, что показывает невысокую развитость отечественных разработок и исследований в данном направлении.

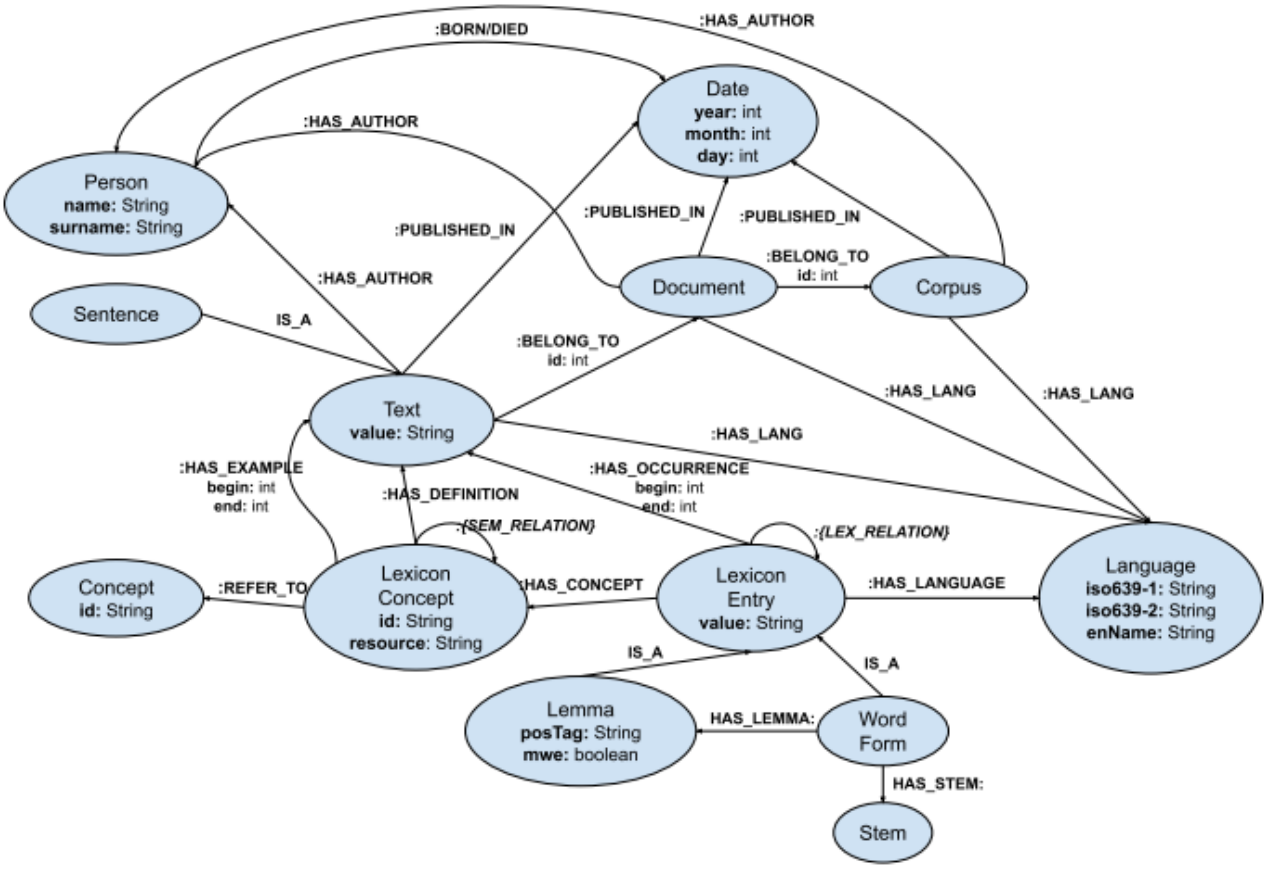

Рис. 1. Модель лингвистического графа знаний

Рассмотрим один из примеров, который, на наш взгляд, наиболее близок к требованиям, предъявляемым к графам знаний для представления тюркского лингвистического ресурса. Таким примером является лингвистический граф, описанный в работе [3] (модель данного лингвистического графа знаний представлена на рисунке 1). По утверждению авторов, этот граф позволяет моделировать:

1. Отношения между концептами и их лексическим представлением;
2. Информацию о статистике слов;
3. Диахроническую информацию как понятий, так и слов.

Лингвистический граф, описанный в названной работе, включает такие вершины, как концепты (Concept), лексические концепты (Lexicon Concept) и лексические входы или лексемы (Lexicon Entry). Лексические концепты связаны между собой таксономическими отношениями типа гипонимии и гиперонимии. Лексические входы связаны как с леммой (Lemma), так и с основой словоформы (Stem). Лемма – это словарная форма. В ряде случаев лемма и основа совпадают, что зависит от типа языка.

Недостатком этого графа знаний для представления полнотекстовой информации является то, что в нем отсутствует возможность описания ситуационно-фреймовой семантики. Данный граф позволяет описывать только лексическую информацию, аналогичную той, что представлена в известном электронном лингвистическом тезаурусе WordNet.

Для описания ситуационных сценариев подходящим ресурсом являются графы знаний фреймового типа. Самыми известными и наиболее заполненными из них являются такие лингвистические базы знаний, как FrameNet и VerbNet. Поэтому необходимо включение в многоуровневый лингвистический граф знаний элементов ресурсов такого типа.

Также при том, что данный граф предназначен для представления словарной информации, в нем также отсутствует возможность грамматической (морфологической и синтаксической) структуры лингвистических единиц. Учитывая богатую морфологическую структуру тюркских языков, можно утверждать, что это является необходимым условием для представления данных о тюркских текстах.

АРХИТЕКТУРА ЛИНГВИСТИЧЕСКОГО ГРАФА ЗНАНИЙ «TURKLANG»

На основе проделанного анализа лингвистических графов знаний нами была создана модель лингвистического графа знаний «TurkLang» для описания тюркских языков. Главное отличие этого графа знаний основано на структурно-функциональных особенностях тюркских языков. В тюркских языках существует четкое деление на структурные компоненты слова, именуемые морфемами. Та-

кое деление позволяет представить морфологическую структуру тюркской словоформы в виде графа, вершинами которого являются морфемы, а ребрами – порядок следования в словоформе.

Модель лингвистического графа знаний для описания тюркских языков «TurkLang» представляет единый граф знаний, который подразделяется на несколько подграфов. Подобное разделение связано с содержимым этих подграфов и с тем, что представляют собой вершины и ребра этих подграфов (схема разделения на подграфы представлена на рис. 2).

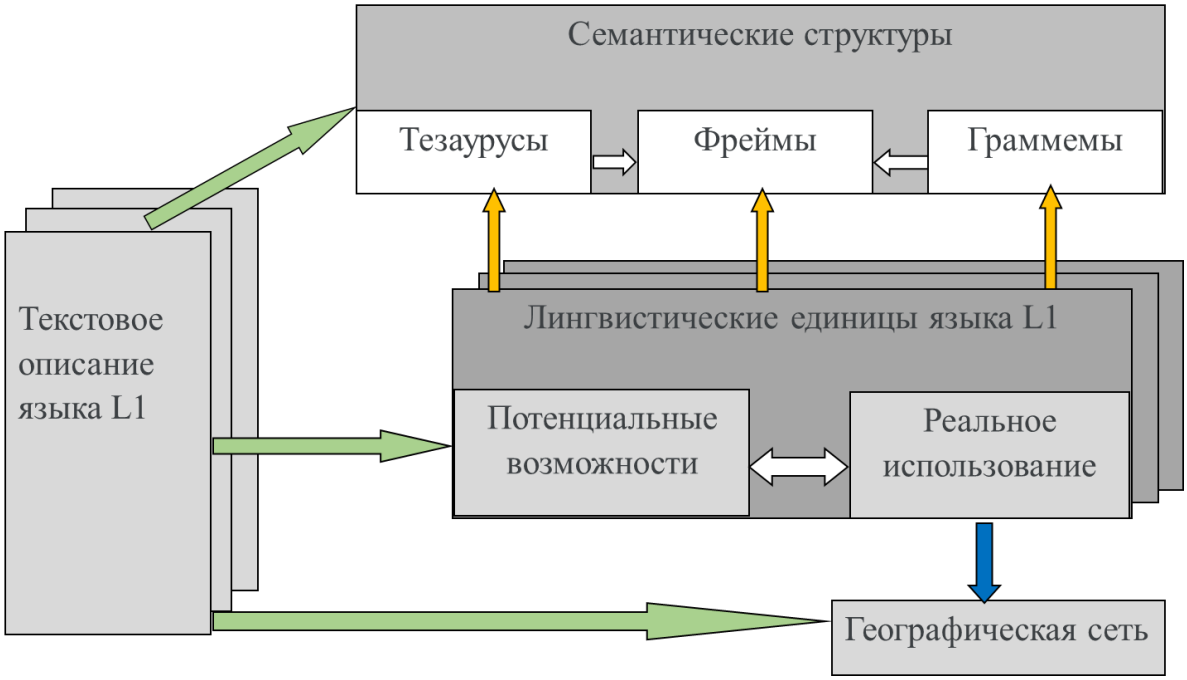

Рис. 2. Архитектура подграфов графа знаний портала

Подграф «Семантические структуры» сам является комбинацией нескольких подграфов, содержащих различные семантические универсалии (семантические единицы, универсальные для всех тюркских языков). «Тезаурусы» – это подграф, вершинами которого являются концепты, связанные между собой отношениями гипонимии и гиперонимии. «Фреймы» – подграф с семантическими сценариями ситуаций, представленными в виде фреймов. Вершинами подграфа «Граммемы» являются грамматические категории и вершины для их классификации.

Подграфы типа «Лингвистические единицы» содержат вершины, соответствующие лингвистическим единицам разных языковых уровней: морфемы, словоформы, аналитические формы, предложения; каждый подграф соответствует

одному тюркскому языку. Ребра отражают структурные связи между этими единицами. На рисунке данный подграф разделен на две составляющие: «Потенциальные возможности» и «Реальное использование», реализующиеся в двух лингвистических ресурсах. «Потенциальные возможности» лингвистических единиц тюркских языков описаны в лингвистическом портале «Тюркская морфема», а «Реальное использование» в речи и тексте представлено в электронном корпусе «Туган тел», который содержит тексты тюркских языков.

Структура модели лингвистического графа знаний «TurkLang» для описания тюркских языков (вершины и связи) и то, из чего состоят подграфы рисунка 2, раскрыта на рис. 3. В центре этого графа изображен элемент «Морфема» (morpheme), который является основной лингвистической единицей графа знаний портала.

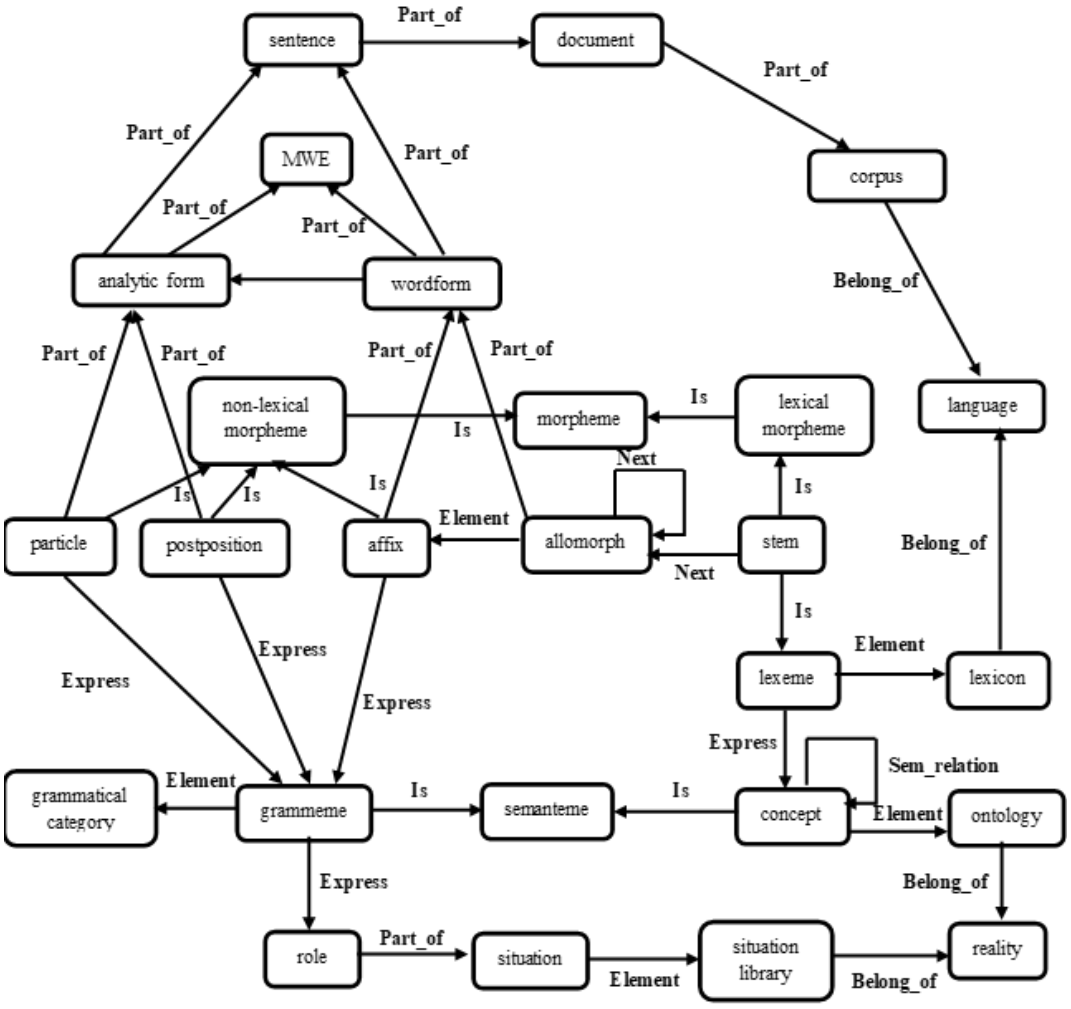

Рис. 3. Структура модели лингвистического графа знаний TurkLang

Вершины, обозначающие «Корпус» (corpus) и «Документ» (document), относятся к разделу графа с описанием реального использования языка. Элементы «Предложение» (sentence), «Морфема» (morpheme), «Корень» (stem), «Аффикс» (affix), «Частица» (particle), «Послелог» (postposition) и «Многословное выражение» (MWE) присутствуют как в разделах реального, так и потенциального описаний языка (см. рис. 2).

Элементы «Грамматическая категория» (grammatical category), «Граммема» (grammeme) и «Семантема» (semanteme) относятся к семантическим структурам грамматики. Элементы «Лексема» (lexeme), «Концепт» (concept) и «Онтология» (ontology) относятся к семантическим структурам тезауруса. Элементы «Ситуация» (situation) и «Роль» (role) относятся к семантическим структурам фреймов. Связь семантических структур с лингвистическими единицами, а также морфотактическая связь лингвистических единиц между собой выражают потенциальные возможности графа знаний по генерации новых текстовых описаний (разработке анализаторов и синтезаторов текстов на различных уровнях: морфологическом, синтаксическом и семантическом).

ИСПОЛЬЗОВАНИЕ ГРАФА ЗНАНИЙ «TURKLANG» ДЛЯ РАЗРАБОТКИ ИНСТРУМЕНТОВ ОБУЧЕНИЯ ТЮРКСКИМ ЯЗЫКАМ

В работе [1] содержится утверждение, что ученик, изучающий языки, должен учиться видеть языковые структуры, искать аналогии и обобщать полученные данные. Кроме того, использование аутентичных материалов делает речь обучающихся более идиоматичной, приближая ее к речи носителей языка. Мы считаем, что все эти возможности в полной мере реализованы в предложенной нами модели, на основе которой построены лингвистический портал «Тюркская морфема» и электронный корпус «Туган тел».

Рассмотрим примеры того, как из базы знаний, построенной на основе нашей модели, можно извлекать информацию о сравнительных особенностях тюркских языков и строить учебные задания для их изучения. Особенно эффективно это может работать для обучения студентов, уже знающих один тюркский язык, другому тюркскому языку на примерах рассмотрения разницы в ситуацион-

ных фреймах двух соответствующих друг другу предложений в языковой паре. Далее представлены примеры учебных заданий для турецко-татарской языковой пары.

1. В зависимости от ролевой схемы глагола выбирается вариант перевода.

o insanı vuruyor *'он убивает человека'* →
PN(o) N(insan)+**ACC(-yI) V(vur)**+PRES(-Iyor) →
PN(ул) N(кеше)+**ACC(-нЫ) V(үтер)**+PRES(-Й) →
ул кешене үтерә
o insana vuruyor *'он ударяет человека'* →
PN(o) N(insan)+**DIR(-yA**) **V(vur)**+PRES(-Iyor) →
PN(ул) N(кеше)+**DIR(-ГА) V(сук)**+PRES(-Й) →
ул кешегә суга

2. Разные ролевые схемы в разных языках.

o bunu Ayşeye sordu *'он спросил это у Айшы'* →
PN(o) N(bu)+ACC(-yI) N(Ayşe)+**DIR(-yA)** V(sor)+PST_DEF(-du) →
PN(ул) N(бу)+ACC(-нЫ) N(Әйшә)+**ABL(-ДАн)** V(сора)+PST_DEF(-ДЫ) →
ул моны Әйшәдән сорады
o işe başlıyor *'он начинает работу'* →
PN(o) N(iş)+**DIR(-yA)** V(başla)+PRES(-Iyor) →
PN(ул) N(эш)+**ACC(-нЫ)** V(башла)+PRES(-Й) →
ул эшне башлый

Такого рода задания могут быть сгенерированы с использованием примеров реальных предложений на том или ином языке из электронного корпуса, семантической разметки на основе ситуационных фреймов портала «Тюркская морфема», морфогенератора предложений портала для перевода и морфоанализатора портала для получения схемы разбора. Также ученику доступна информационно-справочная система портала, с помощью которой возможно более подробное изучение морфем, грамматики и семантики задания.

Ролевые схемы на основе ситуационных фреймов портала «Тюркская морфема» могут быть использованы не только при генерации заданий, но и для автоматизированной проверки ответа ученика. Для этого предлагается следующая реализация прагматически-ориентированного алгоритма автоматического анализа

ответа обучаемого с использованием фреймов, тезауруса и грамматики портала (в частности, морфотактики). Схема этого алгоритма представлена на рис. 4.

Рис. 4. Схема алгоритма анализа ответа

Рассмотрим алгоритм:

1. На первом этапе ответ на вопрос в виде текста на тюркском языке сначала проходит этап морфоанализа с помощью анализатора портала «Тюркская морфема».

2. Далее полученный морфологический разбор поступает на вход лексическому процессору, в котором производятся лексический анализ с использованием модели ответа (свой для каждого вопроса) и трансформация текста ответа в форму цепочки концептов из тезауруса портала (канонизированный ответ). Модель ответа своя для каждого вопроса, она определяет ожидаемый лексикон и семантику ответа, имея вид пар «Концепт – Множество корневых морфем, соответствующих ответу». Такая модель ответа может быть сгенерирована полностью в автоматическом режиме с использованием лингвистического ресурса портала «Тюркская морфема» и ситуационного фрейма (какого – указано далее).

3. На третьем этапе канонизированный ответ поступает на вход семантическому интерпретатору, который производит проверку соответствия цепочки концептов индивидуальной концептуальной грамматике ответа на основе ситуационного фрейма, ожидаемого в ответе для данного задания. В большинстве случаев это тот же ситуационный фрейм, с помощью которого это задание и было сгенерировано. На выходе получается числовой вектор, называемый вектором ситуации. Этот вектор должен позволять оценить правильность, точность и полноту ответа, содержать данные о соответствии ответа ожидаемому ситуационному фрейму, ожидаемому лексикону, о длине ответа, модальность и т. д.

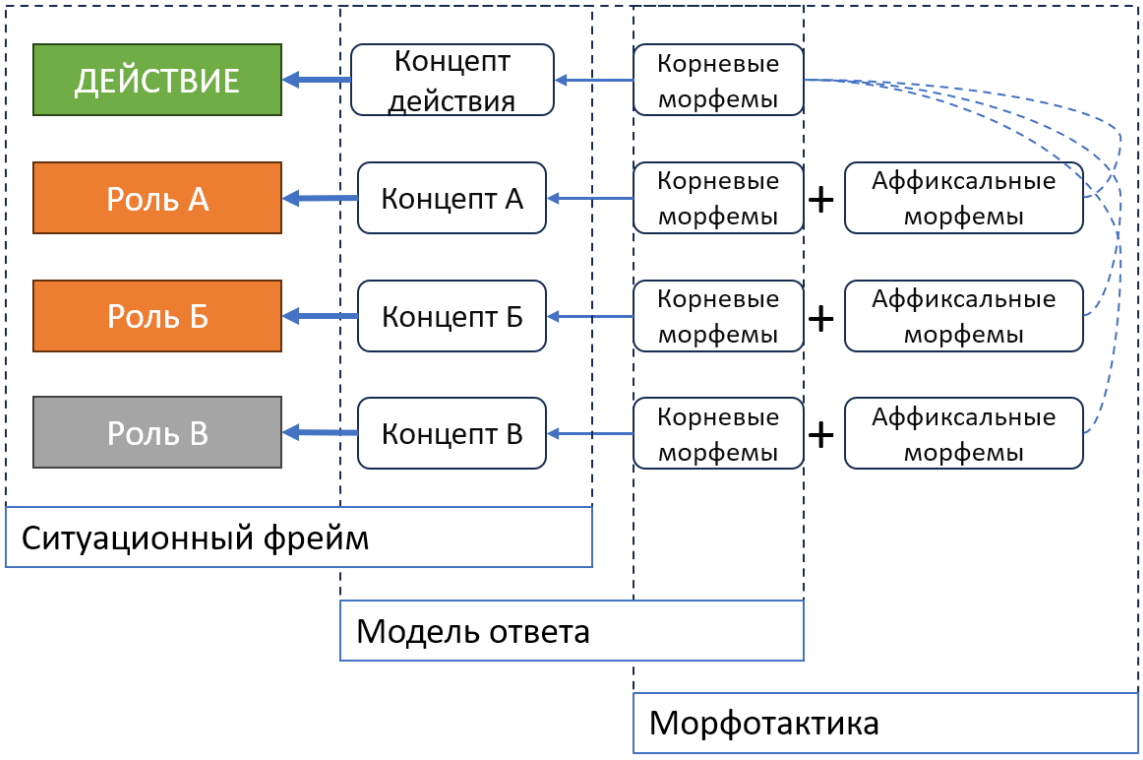

Рис. 5. Схема данных, извлекаемых из портала для анализа ответа

На рис. 5 представлена схема данных, извлекаемых из портала в процессе анализа. Ситуационный фрейм задается действием, определяющим ситуацию и роли объектов, участвующих в данной ситуации. Действию соответствует некоторый набор концептов действия, а ролям объектов – концепты объектов, которые могут выполнять данные роли. К перечисленным концептам могут также относиться атрибутивные единицы – концепты атрибутов действий и концепты атрибутов объектов – но они не обязательны к использованию в фрейме, значит, и в

ответе. Каждому концепту в некотором тюркском языке соответствует множество корневых морфем, и данные пары определяют модель ответа. С корневыми морфемами и между собой правилами морфотактики связаны аффиксальные морфемы. Кроме того, некоторые аффиксальные морфемы обязательны к использованию при реализации ситуационного фрейма в языке. Все эти элементы в полной мере определяют грамматику и семантику всевозможных вариаций правильных ответов (с учетом строго структурированного синтаксиса тюркских языков), что позволяет произвести анализ и предварительную оценку в виде вектора ситуации в автоматическом режиме.

Ранее реализация и оценка аналогичного алгоритма были представлены авторами в статье [4], однако в нем использовались иные языковые универсалии, недостаточно учитывающие особенности тюркских языков и не имеющие ресурсов для автоматической генерации. Представленный здесь алгоритм в полной мере использует ресурс портала «Тюркская морфема» для автоматического анализа ответа. При этом для генерации заданий может быть использован ресурс электронного корпуса «Туган тел» и иных электронных корпусов тюркских языков.

ЗАКЛЮЧЕНИЕ

Представленная модель лингвистического графа «Turklang» находит свою реализацию в разработанных ранее и разрабатываемых на данный момент лингвистических ресурсах и инструментах обработки естественного языка, таких как портал «Тюркская морфема» и электронный корпус тюркских языков «Туган тел». Данный лингвистический граф позволяет наиболее полно представить лингвистическую информацию для тюркских языков, с учетом их структурно-функциональных особенностей, на всех языковых уровнях: грамматика (морфология и синтаксис) и семантика (тезаурус и ситуационные фреймы) как в их потенциальных возможностях, так и в реальном использовании в текстах и речи. За счет этого возможны реализация ресурсов для электронного образования, учебных курсов с использованием информационно-справочной системы, а также разработка автоматических генераторов учебных заданий, внедренных в данные ресурсы и курсы, что является дальнейшей задачей, стоящей перед авторами статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Johns T.* Should you be persuaded. Two samples of data-driven learning materials // T. Johns & P. King (Eds.). Classroom Concordancing. ELR Journal. 1991. № 4. P. 1–16.

2. *Левенкова А.Ю., Трифонова И.С.* Базы данных в лингвистике и языковом образовании: современное состояние и возможности их использования при обучении иностранному языку // Известия ВГПУ. 2023. №2 (175). С. 90–101.

3. *Basile P., Cassotti P., Ferilli S., McGillivray B.* A New Time-sensitive Model of Linguistic Knowledge for Graph Databases // Proc. of the 1st Workshop on Artificial Intelligence for Cultural Heritage co-located with the 21st International Conference of the Italian Association for Artificial Intelligence (AixIA 2022), CEUR Workshop Proceedings. 2022. V. 3286. P. 69–80.

4. *Suleymanov D., Prokopyev N.* Development of Prototype of Natural Language Answer Processor for e-Learning // Kuznetsov S.O., Panov A.I., Yakovlev K.S. (Eds.). Artificial Intelligence. RCAI 2020. Lecture Notes in Computer Science. 2020. V. 12412. P. 448–459.

LINGUISTIC KNOWLEDGE GRAPH “TURKLANG” FOR CREATION OF TOOLS FOR TEACHING TURKIC LANGUAGES

A. R. Gatiatullin[1] [0000-0003-3063-8147], N. A. Prokopyev[2] [0000-0003-0066-7465]

[1]*Tatarstan Academy of Sciences;* [2]*Kazan Federal University*

[1]ayrat.gatiatullin@gmail.com, [2]nikolai.prokopyev@gmail.com

Abstract

This article presents elements of the linguistic knowledge graph “Turklang”, developed at the Institute of Applied Semiotics of the Academy of Sciences of Tatarstan and used as a basis for creating a number of linguistic resources and tools: the portal “Turkic Morpheme”, the electronic corpus of the Tatar language “Tugan Tel”, morpho-analyzer. Creating an educational environment requires subject-oriented knowledge graphs, for which methods of general and open graphs are not suitable. This paper

describes linguistic knowledge graphs, which reflect, on the one hand, potential capabilities of Turkic languages, and on the other hand, examples of actual use in texts. Peculiarity of these knowledge graphs is that they contain linguistic units of different linguistic levels, and concepts corresponding to meanings of these linguistic units, which are built into the thesaurus of concepts. Structure of this knowledge graph allows to formulate the content of a training course, build an individual educational trajectory, as well as create tests and tools of automated answer grading as part of knowledge control when teaching Turkic languages. This makes it possible to subsequently develop, based on these graphs, training programs taking into account the structural and functional features of the Turkic languages, and also contributes to the implementation of individual goals of students.

Keywords*: knowledge graph, knowledge base, linguistic resource, linguistic unit, low-resource languages, Turkic languages, web portal, e-learning, knowledge control, automated answer grading*

REFERENCES

1. *Johns T.* Should you be persuaded. Two samples of data-driven learning materials // T. Johns & P. King (Eds). Classroom Concordancing. ELR Journal. 1991. № 4. P. 1–16.

2. *Levenkova A.Yu., Trifonova I.S.* Bazy dannykh v lingvistike i yazykovom obrazovanii: sovremennoe sostoyanie i vozmozhnosti ikh ispol'zovaniya pri obuchenii inostrannomu yazyku // Izvestiya VGPU. 2023. № 2 (175). P. 90–101.

3. *Basile P., Cassotti P., Ferilli S., McGillivray B.* A New Time-sensitive Model of Linguistic Knowledge for Graph Databases // Proceedings of the 1st Workshop on Artificial Intelligence for Cultural Heritage co-located with the 21st International Conference of the Italian Association for Artificial Intelligence (AixIA 2022), CEUR Workshop Proceedings. 2022. V. 3286. P. 69–80.

4. *Suleymanov D., Prokopyev N.* Development of Prototype of Natural Language Answer Processor for e-Learning // Kuznetsov S.O., Panov A.I., Yakovlev K.S. (Eds). Artificial Intelligence. RCAI 2020. Lecture Notes in Computer Science. 2020. V. 12412. P. 448–459.

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.

Ayrat Rafizovich GATIATULLIN – born in 1972. Graduated from Kazan State University in 1994, candidate in technical sciences (2002). Leading researcher at the Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan. List of scientific works includes more than 60 publications.

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

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 более 40 работ.

Nikolai Arkadievich PROKOPYEV – born in 1992. Graduated from the Institute of Computational Mathematics and Information Technologies of the Kazan Federal University in 2015. Researcher at the Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan. List of scientific works includes more than 40 publications.

email: nikolai.prokopyev@gmail.com
ORCID: 0000-0003-0066-7465;
Author ID (РИНЦ): 999214;
Author ID (Scopus): 57190803409;
Researcher ID (WoS): S-3829-2016.

Материал поступил в редакцию 12 мая 2024 года