II SHO‘BA
TIL TA'LIMIDA KOMPYUTER TEXNOLOGIYALARI

БАЗА ЗНАНИЙ ПОРТАЛА «ТЮРКСКАЯ МОРФЕМА » КАК ЛИНГВИСТИЧЕСКИЙ РЕСУРС ДЛЯ СОЗДАНИЯ ИНСТРУМЕНТОВ ОБУЧЕНИЯ ТЮРКСКИМ ЯЗЫКАМ

Гатиатуллин Айрат Рафизович
Институт прикладной семиотики Академии Наук
Республики Татарстан.

Абдурахмонова Нилуфар
заведующий кафедрой Компьютерной и прикладной лингвистики,
доктор наук, профессор Национального университета Узбекистана.

Аннотация: В статье представлены элементы графов знаний портала «Тюркская морфема», которые представляют собой лингвистический ресурс для разработки учебных курсов по изучению тюркских языков. Сам интернет-портал «Тюркская морфема» – это web-сайт (modmorph.turklang.net), который включает набор различных сервисов на базе лингвистических ресурсов по тюркским языкам, и ориентирован на работу с тюркскими языками в разных аспектах: морфонологическом, морфологическом, синтаксическом, семантическом. Создание образовательной среды требует предметно-ориентированные графы знаний, для получения которых не подходят методы создания общих и открытых графов. В данной работе описываются лингвистические графы знаний, которые, с одной стороны, отображают потенциальные возможности тюркских языков, а с другой стороны, примеры реального использования в текстах на тюркских языках. Особенность этих графов знаний в том, что, с одной стороны, они содержат лингвистические единицы разного языкового уровня, а с другой стороны – концепты, соответствующие значениям этих лингвистических единиц, которые встроены в тезаурус концептов. Структура такого графа знаний позволяет формировать контент учебного курса, строить индивидуальную образовательную траекторию, а также формировать задания и средства автоматизированной их 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.

1.Введение

В настоящее время наблюдается бурное развитие систем автоматической обработки естественных языков (распознавания речи, автоматического перевода, анализа тональности,), которые основаны на технологиях машинного обучения. Данные технологии требуют большого количества ресурсов для эффективной работы. Однако языков с достаточной степенью изученности и готовыми большими наборами данных сравнительно мало. В список высокоресурсных языков входят такие языки, как английский, китайский (мандарин), испанский, французский, немецкий, японский, и несколько других языков.

Среди тюркских языков в данный список можно включить только турецкий язык, хотя оценка специалистов по степени оснащенности ресурсами турецкого языка неоднозначная. Одни авторы, называют его малоресурсным языком, а обзор сделанный в работе [1], показывает наличие большого количества лингвистических ресурсов для турецкого языка.

Сам термин малоресурсные языки был введен еще в 2003 год Краувером [2].

Согласно его определению малоресурсные языки – это естественные языки, обладающие следующими свойствами:

1. недостаток своей системы письменности или устойчивой орфографии;
2. нехватка квалифицированных лингвистов и переводчиков для данного языка;
3. ограниченное распространение в сети Интернет;
4. нехватка электронных ресурсов для обработки языка и речи, в том числе одноязычных корпусов, двуязычных электронных словарей, орфографических и фонетических транскрипций речи, словарей произношения и т. д.

Низкий уровень развития речевых технологий для малоресурсных языков объясняется рядом причин.

1. Задача описания и формализации структур малоресурсного языка достаточно трудоемка с научной точки зрения и сопряжена с высокими временными, людскими и финансовыми затратами для подготовки необходимых ресурсов (речевых корпусов, программного обеспечения).
2. Существующие на сегодня способы моделирования и создания программ для обработки речи направлены на решение узкого круга задач и не учитывают особенности работы с малоресурсными языками.

Подготовка многоязычных ресурсов требует разного уровня сопоставления или выравнивания лингвистических ресурсов разного уровня. В настоящее время можно выделить три уровня выравнивания лингвистических ресурсов на разных языках: документ, предложение и слово. Ясно, что выстраивание соответствий между этими языковыми единицами приводит к увеличению объема информации (в указанном порядке) за счет добавления новых структурных элементов. Полученные структуры с выстроенными соответствиями представляют собой графы знаний в узлах которых может храниться информация разного уровня в зависимости от уровней выравнивания. При выравнивании документов в узлах графа знаий могут храниться целые документы, при выравнивании на уровне предложений узлами графа могут быть предложения. При выравнивании на уровне более мелких лингвистических единиц соответствия не всегда выстраиваются между отдельными словами, в некоторых случаях отдельному слову может соответствовать аналитическая форма или аналитическая конструкция.

Разные типы соответствий используются для решения разного типа задач. Так выравнивание на уровне документа используется для поиска информации. Выравнивание на уровне предложения используется в задачах машинного перевода. Выравнивать можно, как неструктурированные, так и структурированные предложения представленные в виде синтаксических деревьев или семантических фреймов[Abdurakhmonova, N. 2019,2021,2022].

Выравнивание на уровне слов или лексем лежит в основе большого количества лингвистических задач.

Базы знаний, полученные в виде графов знаний, с выровненными лингвистическими единицами имеют достаточно широкий спектр применения, к примеру, они являются необходимым компонентом при разработке обучающих систем иностранным языкам. Для обучения тюркским языкам необходимы ресурсы, содержащие тюркские лингвистические базы знаний, что позволяет использовать такие базы знаний для создания многоязычных обучающих систем при обучении одному из тюркских языков. Например, это может быть обучение одному из тюркских языков лиц, владеющих другими тюркскими языками. Это достаточно актуальная задача поскольку существующие тюркские языки зачастую достаточно сильно отличаются друг от друга, и носители одного тюркского языка не всегда понимают носителя другого тюркского языка.

В данной работе рассматривается лингвистическая база данных в виде графа знаний портала «Тюркская морфема» (modmorph.turklang.net) [3] с точки зрения потенциала его использования в задачах образования.

2. Графы знаний

В последнее десятилетие, одним из эффективных способов представления лингвистической информации в ресурсах являются графы знаний. Имеется целый ряд работ с описанием лингвистических графов 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 [4] (модель данного лингвистического графа знаний представлена на рисунке 1). По утверждению авторов, данный граф позволяет моделировать:

1) Отношения между понятиями и словами;
2) Информацию о встречаемости слов;
3) Диахроническую информацию как понятий, так и слов.

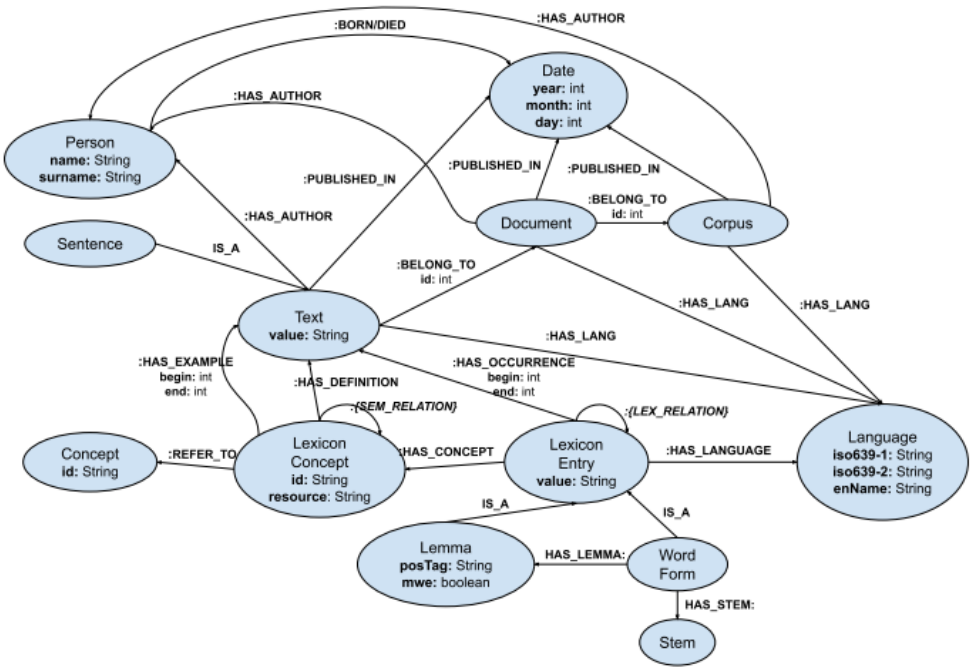

Рисунок 1. Модель лингвистического графа знаний

Описанный в данной работе лингвистический граф включает такие вершины, как Концепты (Concept), Лексические концепты (Lexicon Concept), Лексемы (Lexicon Entry). Лексические концепты взаимосвязаны между собой семантическими отношениями типа гипонимии и гиперонимии. Лексемы взаимосвязаны с леммой (Lemma) и основой словоформы (Stem). Особенность данного графа знаний в том, что он не описывает ситуационную семантику, поэтому его семантическое описание является достаточно ограниченным и отражает только таксономические отношения, аналогичные тем, что представлены в известном ресурсе WordNet. Для описания ситуационной семантики подходят графы знаний фреймового типа, например такой достаточно известный ресурс, как FrameNet, отсюда возникла идея добавления ситуационных фреймов к структуре данного лингвистического графа при реализации ресурса портала «Тюркская морфема».

Также в данном графе знаний не представлена грамматическая (морфологическая) структура словоформ, а для тюркских языков это является необходимым свойством, так как данные языки обладают богатой морфологией агглютинативного типа. В отличие от языков флективного типа, к которым относится русский язык, в тюркских языках существует четкое деление на структурные компоненты слова, которые называются морфемами. Такое деление позволяет представить морфологическую структуру словоформы в виде подграфа, вершинами которого являются морфемы, что также реализовано в графе знаний портала «Тюркская морфема».

3. Архитектура лингвистической базы знаний портала

Лингвистическая база данных портала «Тюркская морфема» представляет собой единый граф знаний, который подразделяется на несколько подграфов. Разделение на подграфы сделано в связи со структурными особенностями каждого из этих подграфов, а также с тем, что каждый из подграфов содержит наборы вершин одного типа. Вершины одного подграфа связаны между собой отношениями одного типа, а с вершинами из других подграфов отношениями иного типа. Схема разделения на подграфы представлена на рисунке 2. Такое разделение связано и с задачами, для решения которых используются каждый из подграфов единого графа знаний портала. Далее рассмотрим подграфы знаний портала, объединяемые в единый граф знаний.

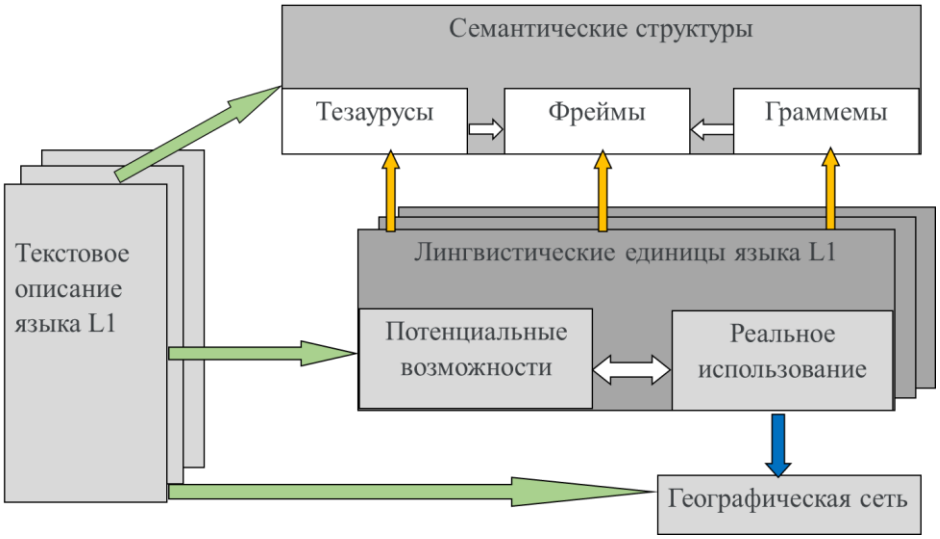

Рисунок 2. Архитектура подграфов графа знаний портала

Элементы всех подграфов портала «Тюркская морфема» объединяются в единый граф знаний, схема которого представлена на рисунке 3. В центре данного графа изображен элемент «Морфема» (morpheme), который является основной лингвистической единицей графа знаний портала.

Элементы «Корпус» (corpus), «Документ» (document), «Предложение» (sentence), относятся к текстовому описанию языка (см. рисунок 2) и указывают на реальное использование лингвистических единиц в текстах на некотором тюркском языке.

Элементы «Морфема» (morpheme), «Корень» (stem), «Аффикс» (affix), «Частица» (particle), «Послелог» (postposition), «Многословное выражение» (MWE) являются лингвистическими единицами языка, а их связи с текстовым описанием языка отображают реальное использование [Abdurakhmonova, N. 2019,2021,2022].

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

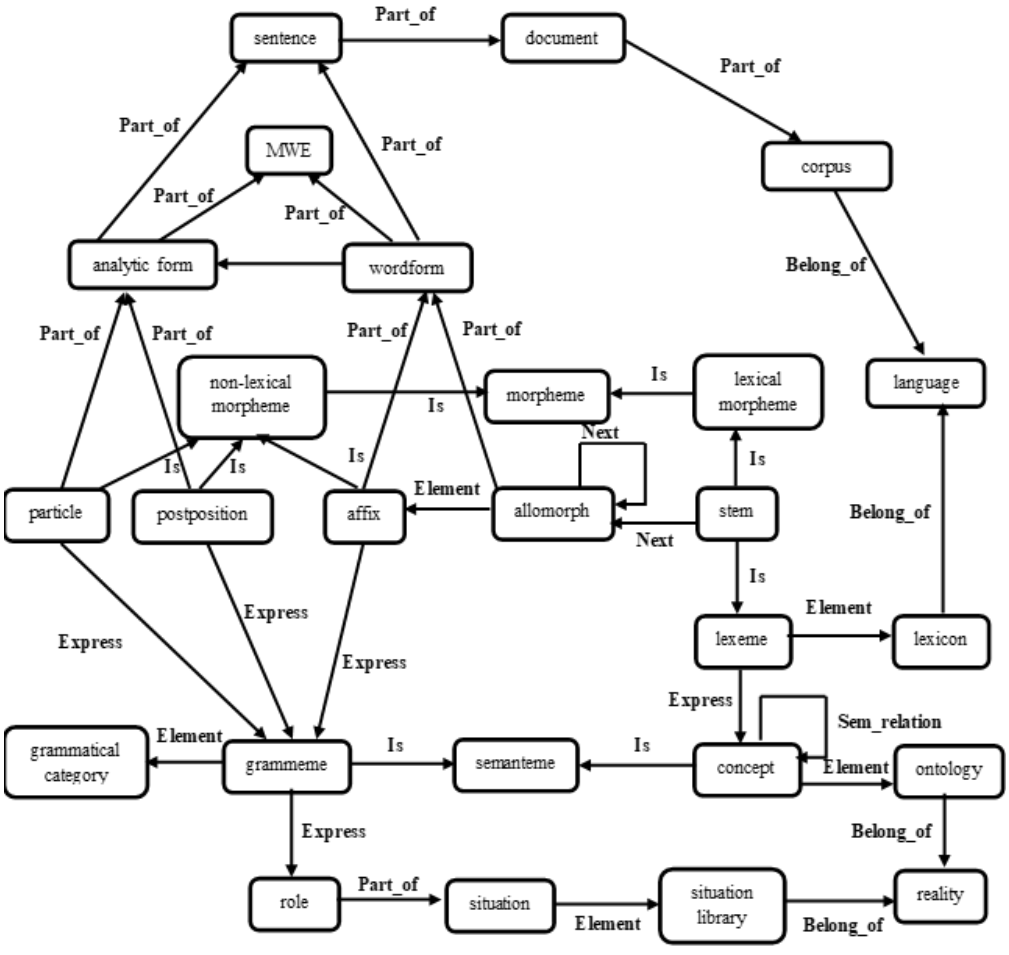

Рисунок 3. Схема графа знаний портала «Тюркская морфема»

4. Использование графов знаний портала в процессе обучения.

Одним из основных блоков графа знаний портала является подграф для описания ситуационных фреймов. Ситуационные фреймы представляют собой описание типовых ситуаций в языке, которые представляют собой более 1000 различных фреймов. Для каждой ткой ситуации дается текстовое описание ситуации (Рис.4) и наборы ролей типового фрейма (Рис.5).

Motion : Двигаться

Название (Русский)	Двигаться
Описание (Русский)	Некоторая сущность (Тема) начинается в одном месте (Источник) и заканчи месте (Цель), покрыв некоторое пространство между ними (Путь). В качест указать Область или Направление, в котором движется Тема, или Расстоян
Название (Английский)	Motion
Описание (Английский)	Some entity (Theme) starts out in one place (Source) and ends up in some other space between the two (Path). Alternatively, the Area or Directionin which the The movement may be mentioned.

Рисунок 4. Описание типовой ситуации

Роли в фреймах

Роль	Описание (Русский)	Описание (Английский)
Source : Источник	Источник-это место, которое Тема занимает изначально до ее изменения местоположения.	The Source is the location the Theme occupies initially before its change of location.
Theme : Тема	Тема-это объект, который меняет местоположение. Обратите внимание, что это не обязательно самодвижущийся двигатель.	The Theme is the entity that changes location. Note that it is not necessarily a self-mover.
Direction : Направление	Это ФЕ используется для выражений, которые указывают на движение по линии от центра божества к опорной точке (которая	This FE is used for expressions that indicate motion along a line from the deitic center towards a reference point (which may be implicit) that is neither

Рисунок 5. Набор ролей типового фрейма

За основу базы знаний портала “Тюркская морфема” взяты набор типовых ситуационных фреймов FrameNet и наборы типовых ролей этого ресурса. Все семантические роли выражаются грамматическими элементами: аффиксальными или аналитическими средствами (Рис.6). Для каждого языка будет свой набор грамматических форм, выражающих определенные семантические роли.

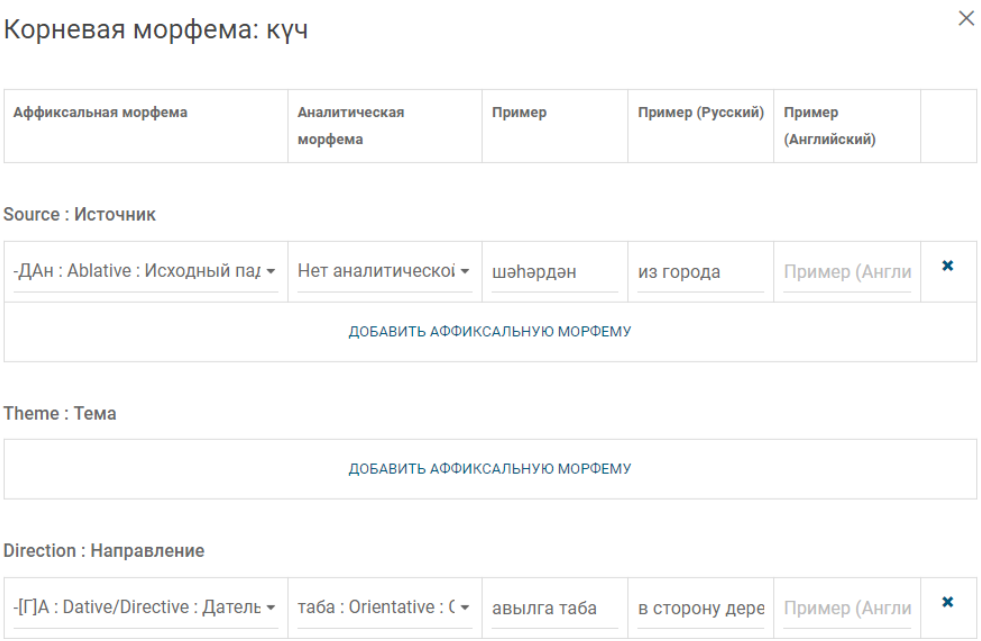

Рисунок 6. Грамматическое выражение типовых ролей фрейма

Для каждого из тюркских языков будет свой набор глагольного (ситуационного) управления.

Рассмотрим прямое использование графа знаний как информационно-справочной системы для составления заданий при обучении некоторому тюркскому языку студентов, уже знающих один тюркский язык на примере рассмотрения разницы в ситуационных фреймах татарских и турецких глаголов с использованием морфогенератора предложений.

1. В зависимости от ролевой схемы глагола выбирается вариант перевода
vurmak – убить / ударить

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

Заключение

В данной статье представлен очередной способ использования фреймового блока портала “Тюркская морфема”.

Список литературы:

1. Çöltekin Ç, Doğruöz AS, Çetinoğlu Ö. Resources for Turkish natural language processing: A critical survey. Lang Resour Eval. 2023;57(1). –P. 449-488. http://doi.org/10.1007/s10579-022-09605-4 Epub 2022 Aug 26. PMID: 36060268; PMCID: PMC9417072.
2. Krauwer S. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. Proceedings of International workshop on speech and computer (SPECOM-2003). 2003. -P. 8–15.
3. Abdurakhmonova, N. Z., Ismailov, A. S., & Mengliev, D. (2022, November). Developing NLP Tool for Linguistic Analysis of Turkic Languages. In 2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON) (pp. 1790-1793). IEEE.
4. Abdurakhmonova, N., Alisher, I., & Toirova, G. (2022, September). Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language Processing. In 2022 7th International Conference on Computer Science and Engineering (UBMK) (pp. 73-75). IEEE.
5. Gatiatullin, A., Suleymanov, D., Prokopyev, N., & Abdurakhmonova, N. (2022, November). “Turkic Morpheme”: From the Portal to the Linguistic Platform. In World Conference Intelligent System for Industrial Automation (pp. 181-189). Cham: Springer Nature Switzerland.
6. Abdurakhmonova, N. (2019). Dependency parsing based on Uzbek Corpus. In of the International Conference on Language Technologies for All (LT4All).
7. Agostini, A., Usmanov, T., Khamdamov, U., Abdurakhmonova, N., & Mamasaidov, M. (2021, January). Uzwordnet: A lexical-semantic database for the uzbek language. In Proceedings of the 11th Global Wordnet conference (pp. 8-19).
8. Gatiatullin A., Kubedinova L., Prokopyev N, Suleymanov D. Linguistic knowledge graphs of the “Turkic morpheme” portal // 2023 8th International Conference on Computer Science and Engineering (UBMK), 2023. –P. 408-413
9. Basile P., Cassotti P., Ferilli S., McGillivray B. 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. –P. 3286.