Linguistic Knowledge Graph "Turklang" as Universal Model for Linguistic Resources and Tools in Turkic Languages

Ayrat Gatiatullin
Institute of Applied Semiotics Tatarstan Academy of Sciences
Kazan, Russia
ayrat.gatiatulin@gmail.com

Nikolai Prokopyev
Institute of Computational Mathematics and Information Technologies Kazan Federal University
Kazan, Russia
nikolai.prokopyev@gmail.com

Lenara Kubedinova
Institute of Applied Semiotics Tatarstan Academy of Sciences
Kazan, Russia
kubedinova@gmail.com

Nilufar Abdurakhmonova
National Univ. of Uzbekistan Dept. Computational and Applied Linguistics
Tashkent, Uzbekistan
n.abduraxmonova@nuu.uz

Rustam Burnashev
Institute of Computational Mathematics and Information Technologies Kazan Federal University
Kazan, Russia
r.burnashev@inbox.ru

2024 9th International Conference on Computer Science and Engineering (UBMK) | 979-8-3503-6588-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/UBMK63289.2024.10773543

***Abstract*—This paper describes the model linguistic knowledge graph "Turklang", as a kind of universal for creating tools for computer processing of Turkic languages. This universal is currently being the basis of linguistic portal "Turkic Morpheme", developed at the Institute of Applied Semiotics of Tatarstan Academy of Sciences. The portal is a multifunctional linguistic toolkit, and one of its functions is accumulation of data on the potential capabilities of Turkic languages for further use in information-reference and educational systems. These potential capabilities of Turkic languages form a whole global potential knowledge graph, while their actual use in real texts forms only a subgraph of this graph, due to the fact that native speakers do not use all the capabilities of their language, either spelling shortened speech act, or not knowing the language fully. A certain group of subgraphs in the global potential knowledge graph "Turklang" will be subject-oriented knowledge graphs, with the help of which individual educational trajectories will be formed and automated assessment tools will be created as part of knowledge control in teaching Turkic languages.**

Keywords— knowledge graph, linguistic resource, linguistic unit, low-resource languages, Turkic languages

I. Introduction

In 1991, T. Jones [1] put forward the idea of teaching languages based on linguistic databases. He made a hypothesis that teaching languages would be more effective if the learner himself acts as a language researcher, and the teacher provides him with context and directions for language mastering. Thus, learners have an opportunity to work with linguistic databases and conduct research for their educational purposes. Jones highlights the following advantages of this approach to teaching:

1) The learner start seeing the language structures, look for analogies and generalize the obtained data. In addition, use of authentic materials makes his speech more idiomatic, bringing it closer to native speech.

2) The teacher, instead of mediating the information about the language, becomes a coordinator of student's research, which allows the learner to independently process information "about the ways of using a language form, as well as solve problems related to its understanding".

3) Role of grammar in learning a foreign language changes. It is argued that language grammar isn't able to reflect the full diversity of its syntactic structures, so it is ineffective to study grammar separately from the area of real functioning of grammar rules. Use of linguistic resources in the educational process can make this process more natural.

We believe that these statements are also true for Turkic languages, which have rich morphology, and to test this hypothesis, it is necessary to have linguistic resources containing Turkic linguistic knowledge bases. Given the structural similarity, these resources can be universal for all Turkic languages. Multilingual nature of these resources can play a positive role in that the learner will be able to compare language examples in different Turkic languages, focusing on the features of his studied language. In the work [2] authors use the ideas put forward by T. Jones [1] and identify 3 types of educational materials and methods that students can use to study a foreign language:

1) Electronic lexicography;
2) Corpus studies;
3) Designing the electronic textbooks and terminology databases.

To the advantages of studying a foreign language using these materials and methods, taking into account the ideas of T. Jones, the authors of [2] added a number of additional points:

1) Working with relevant material;
2) Developing research skills;
3) Mastering the natural structure of speech.

They also believe that the value of a linguistic database in studying foreign languages increases if it provides educational, reference, systematizing, and communicative functions.

We propose to test all these the hypotheses on Turkic languages case study. To verify these statements, it is necessary to have linguistic resources, which, as already stated, should provide educational, reference, systematizing, and communicative functions. However, at present, almost all Turkic languages, except for Turkish, are low-resource

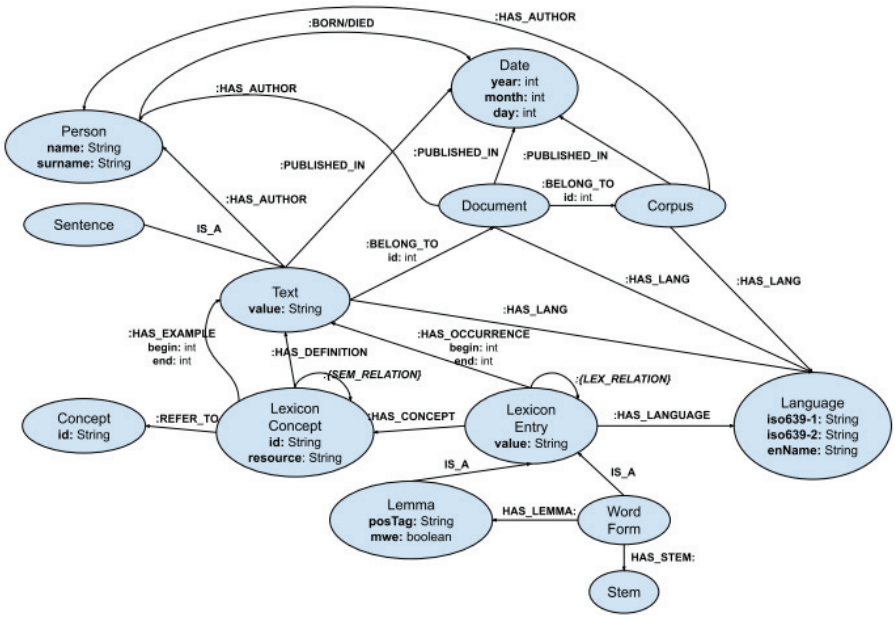

Fig. 1. Model of linguistic knowledge graph

languages, since they lack linguistic resources of various types. A number of linguistic resources for computer processing of Turkic languages are being developed at the Institute of Applied Semiotics of Tatarstan Academy of Sciences. Two of them can be singled out as the most significant:

1) Linguistic portal "Turkic Morpheme";
2) Electronic corpus "Tugan Tel".

The electronic corpus "Tugan Tel" was previously implemented in two versions on different technological platforms, and currently a third version is being developed based on graph databases, as they allow more efficient representation of syntactic and semantic information in the corpus. The new version of "Tugan Tel" and the linguistic portal "Turkic Morpheme" will be based on a unified linguistic knowledge model "Turklang", implemented in form of linguistic knowledge graph. This model is being in development in the Institute of Applied Semiotics of Tatarstan Academy of Sciences.

II. Linguistic Knowledge Graphs

Currently, one of the effective ways of representing linguistic information in resources is knowledge graphs. There are a number of works describing linguistic knowledge graphs, however, mostly for Indo-European languages.

Let us consider one of the examples, which, in our opinion, is closest to the requirements for knowledge graphs suitable for a Turkic linguistic resource. Such an example is the linguistic knowledge graph described in paper [3] (model of this graph is shown in Fig. 1). According to the authors, this graph allows modeling:

1) Relations between concepts and their lexical representation;
2) Information on word statistics;
3) Diachronic information of both concepts and words.

This linguistic graph includes such node types as Concept, Lexicon Concept, Lexicon Entry. Lexicon concepts are linked to each other by taxonomic relations of hyponymy and hypernymy types. Lexicon entries are linked to both Lemmas and Stems, where Lemma is a dictionary form of word and Stem is a morphological base of word. In some cases, Lemma and Stem coincide, depending on language.

Disadvantage of this knowledge graph for representing full-text information is that it does not have the ability to describe situational-frame semantics. This graph allows to describe only lexical information similar to that presented in the well-known electronic thesaurus WordNet. Frame-type knowledge graphs are a suitable resource for describing situational scenarios. The most famous and most full linguistic knowledge bases are FrameNet and VerbNet. Therefore, it is necessary to include elements of resources of this type in a multi-level linguistic knowledge graph. Also, despite the fact that this graph is intended to represent dictionary information, it also lacks the ability to express the grammatical (morphological and syntactic) structure of linguistic units. Considering rich morphological structure of Turkic languages, this is a necessary requirement for representing data on Turkic texts.

III. Architecture of "Turklang" Knowledge Graph

Based on analysis of linguistic knowledge graphs, we created a model of the linguistic knowledge graph "Turklang" for Turkic languages. The main difference of this knowledge graph is based on structural and functional features of Turkic languages. In Turkic languages, there is a clear division into structural components of the word, called morphemes. This division allows us to represent the

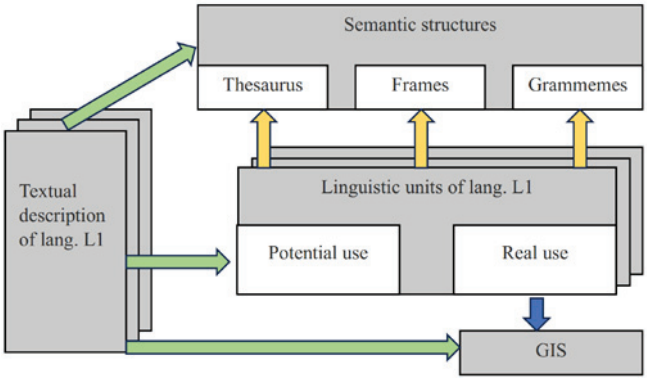

Fig. 2. Architecture of "Turklang" knowledge graph

morphological structure of a Turkic wordform as a graph, the vertices of which are morphemes, and the edges provide their order in the wordform.

Linguistic knowledge graph "Turklang" is architecturally divided into several subgraphs. Such a division is explained by what the vertices and edges of these subgraphs represent (architecture is shown in Fig. 2).

"Semantic structures" subgraph is itself a combination of several components containing various semantic universals (semantic units that are universal for all Turkic languages). "Thesauri" is a component whose vertices are concepts related to each other by hyponymy and hypernymy. "Frames" is a component with semantic scenarios of situations presented in the form of situational frames. Vertices of "Grammemes" component are grammatical categories organized in hierarchy.

Subgraphs of "Linguistic units" type contain vertices corresponding to linguistic units of different language levels: morphemes, word forms, analytical forms, sentences; each subgraph corresponds to one Turkic language. Their edges reflect structural connections between these units. In the figure, this subgraph is divided into two components: "Potential capabilities" and "Actual use", which are realized in two linguistic resources. "Potential capabilities" of linguistic units of Turkic languages are described in the "Turkic Morpheme" portal and "Actual use" in speech and text is presented in the "Tugan Tel" corpus.

Structure of the "Turklang" linguistic knowledge graph model, as what the subgraphs of Fig. 2 consist of, is revealed in Fig. 3 in form of node types and connection types. In the center "Morpheme" is shown, which is the main linguistic unit of "Turklang" knowledge graph.

Nodes denoting "Corpus", "Document" refer to the "Actual use" component of "Linguistic units" subgraphs. The elements "Sentence", "Morpheme", "Stem", "Affix", "Particle", "Postposition", "Multi-Word Expression" (MWE) are present in both "Actual use" and "Potential capabilities" components (see Fig. 2). The elements "Grammatical category", "Grammeme", "Semanteme" refer to "Grammemes" component of "Semantic structures" subgraph. The elements "Lexeme", "Concept", "Ontology" refer "Thesauri" component. The elements "Situation", "Role" refer to "Frames" component. Connection of semantic structures with linguistic units, as well as the morphotactic connection of linguistic units among themselves, expresses the potential capabilities of the knowledge graph to generate new text descriptions used for development of analyzers and synthesizers of texts at various levels: morphological, syntactic, semantic.

IV. Usage of "Turklang" knowledge graph in development of educational tools for Turkic languages

In the work of T. Jones [1] it is stated that a learner studying languages should learn to see language structures, look for analogies and generalize the obtained data. In addition, use of authentic materials makes the speech of learners more idiomatic, bringing it closer to the speech of native speakers. We believe that all these possibilities are fully realized in the knowledge graph model we proposed, on the basis of which the linguistic portal "Turkic Morpheme" and the electronic corpus "Tugan Tel" were built.

Let us consider examples of how data on the comparative

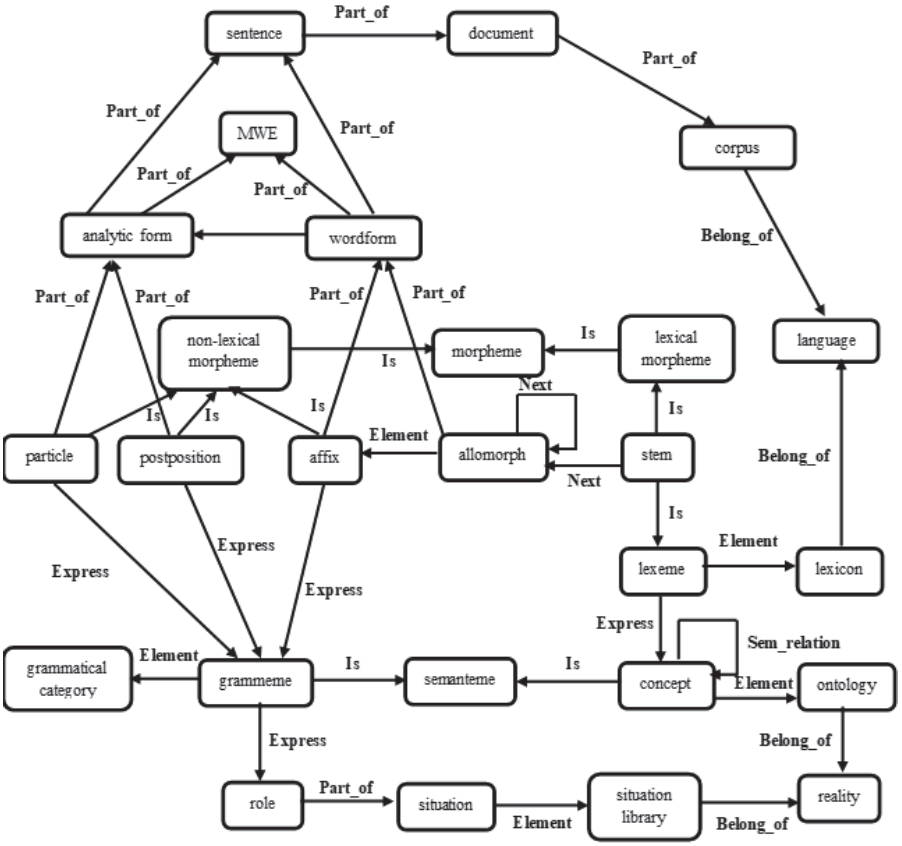

Fig. 3. Structure of "Turklang" knowledge graph

features of Turkic languages can be extracted from the knowledge base built on the basis of our model and how to develop learning tasks using this data. This can work especially effectively for teaching students who already know one Turkic language to another Turkic language using examples of considering the difference in situational frames of two corresponding sentences in a language pair. Below are examples of learning tasks for the Turkish-Tatar pair.

1) Depending on role scheme of the verb, translation option is selected.

o insanı vuruyor *'he kills a person'* → 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 *'he hits a person'* → PN(o) N(insan)+**DIR(-yA) V(vur)**+PRES(-Iyor) → PN(ул) N(кеше)+**DIR(-ГА) V(сук)**+PRES(-Й) → ул кешегә суга

2) Different role schemes in different languages.

o bunu Ayşeye sordu *'he asked Aisha this'* → 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 *'he starts work'* → PN(o) N(iş)+**DIR(-yA)** V(başla)+PRES(-Iyor) → PN(ul) N(esh)+**ACC(-нЫ)** V(башла)+PRES(-Й) → ул эшне башлый

Such tasks can be generated using examples of real sentences in a particular language from the electronic corpus, semantic markup based on situational frames of the portal "Turkic Morpheme", a morpho-generator of portal sentences for translation, and a morpho-analyzer of the portal to obtain a parsing scheme. Also, the student has access to information and reference system of the portal, with help of which a more detailed analysis of morphemes, grammar and semantics from the task is possible.

Role schemes based on situational frames of the portal "Turkic Morpheme" can be used not only for generating tasks, but also for automated assessment of the student's answer. For this purpose, the following implementation of a pragmatically oriented algorithm for automatic processing of the student's answer using frames, thesaurus, grammar and morphotactic rules from the portal is proposed. The diagram of this algorithm is given in Fig. 4.

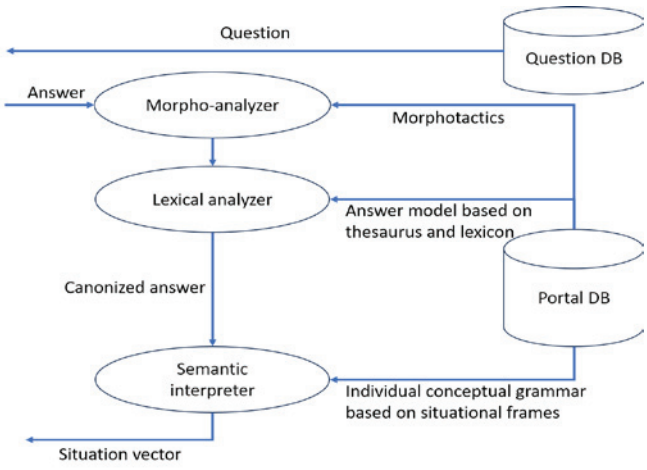

Fig. 4. Scheme of answer processing algorithm

Let's consider the algorithm:

1) At the first stage, the answer to the question in form of a text in Turkic language first undergoes a morphoanalysis using analyzer from the "Turkic Morpheme" portal.

2) Then the obtained morphological analysis result is sent to input of the lexical processor, in which lexical analysis is performed using the answer model and transformation of the answer text into the form of concept chain using the portal thesaurus (this chain is called canonized answer). Answer model is separate for each question, it determines the expected lexicon and semantics of the answer, having the form of pairs "Concept - Set of root morphemes corresponding to the answer". Answer model can be generated completely automatically using the linguistic resource of the "Turkic Morpheme" portal using a situational frame (which one to use is indicated below).

3) At the third stage, the canonized answer is put to the semantic interpreter, which checks correspondence of concept chain to individual conceptual grammar of the answer based on situational frame expected in the answer for this question type. In most cases, this is the same situational frame with which the question was generated. Output is a numerical vector called situation vector. This vector should allow to evaluate the correctness, accuracy, and completeness of the answer, contain data on correspondence of the answer to the expected situational frame and, the on length of the answer, modality, etc.

Fig. 5 shows the diagram of data extracted from the portal during the analysis. Situational frame is specified by action that defines the situation and roles of the objects participating in this situation. Action corresponds to a certain set of action concepts, and roles of objects correspond to the concepts of objects that can perform these roles. The listed concepts may also include attributive units – concepts of action attributes and concepts of object attributes – but they are not required to be used in the frame, and, therefore, in the answer. Each concept in a certain Turkic language corresponds to a set of root morphemes, and these pairs determine the answer model. Affixal morphemes are associated with root morphemes and with each other by the morphotactics rules. In addition, some affixal morphemes are required to be used when implementing a situational frame in the language. All these elements fully determine the grammar and semantics of all possible variations of correct answers, taking into account the strictly structured syntax of

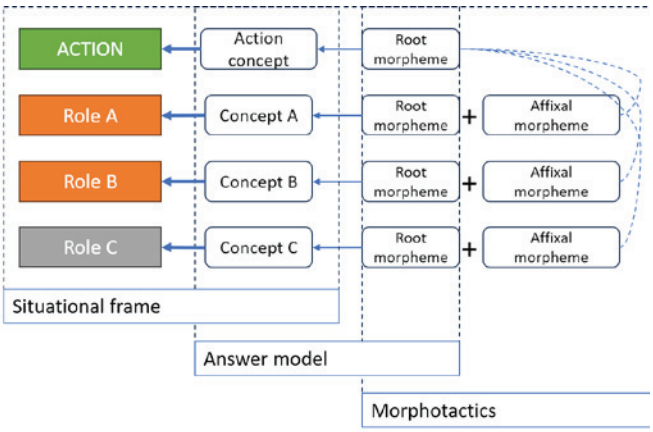

Fig. 5. Data extracted from the portal for answer processing

Turkic languages, which allows for processing and preliminary assessment in form of a situation vector in automatic mode.

Previously, implementation and evaluation of a similar algorithm was presented by the authors in the paper [4], but it used other language universals that did not sufficiently take into account the features of Turkic languages and did not have the resources for automatic generation. The algorithm presented here fully utilizes the resource of "Turkic Morpheme" portal for automatic answer processing. At the same time, resource of the electronic corpus "Tugan Tel" and other electronic corpora of Turkic languages can be used to generate questions.

V. Conclusion

Presented model of the "Turklang" linguistic graph is implemented in previously developed and currently being developed linguistic resources and tools for natural language processing, such as the portal "Turkic Morpheme" and the electronic corpus of Turkic languages "Tugan Tel". This linguistic graph allows the most complete presentation of linguistic information for Turkic languages, taking into account their structural and functional features, at all language levels: grammar (morphology and syntax), semantics (thesaurus and situational frames) both in their potential capabilities and in real use in texts and speech. Due to this, it is possible to implement resources for e-learning, training courses using the information and reference system, as well as to develop automatic generators and assessors of educational tasks embedded in these resources and courses, which is the next course of action in our research.

References

[1] T. Johns, “Should you be persuaded. Two samples of data-driven learning materials,” Classroom Concordancing. ELR Journal, No. 4, 1991, pp. 1-16.

[2] A. Y. Levenkova and I. S. Trifonova, “The data base in the linguistics and linguistic education: the modern state and potential of its usage in the process of the foreign language teaching,” Ivzestia of the Volgograd State PedagogicalUniversity, vol. 175, no. 2, 2023, pp. 90-101.

[3] P. Basile, P. Cassotti, S. Ferilli, and B. McGillivray, “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, Vol. 3286, 2022, pp. 69-80.

[4] D. Suleymanov and N. Prokopyev, “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, vol. 12412, 2020, pp. 448-459.