On Developing of the FrameNet-Like Resource for Tatar

Ayrat Gatiatullin, Alexander Kirillovich and Olga Nevzorova

Kazan Federal University, Kazan, Russia
ayrat.gatiatullin@gmail.com, alik.kirillovich@gmail.com,
onevzoro@gmail.com

Abstract. In this paper, we present TatVerbBank, the first FrameNet-like resource for Tatar language. TatVerbBank is organized as a collection of semantic and syntactic frames. A semantic frame contains semantic roles associated with a concept (for example, for the concept of gift, the roles are giver, recipient, gift, time, etc.). A syntactic frame contains a subcategorization model for a particular Tatar lexical entry and its mapping to semantic roles. The developed resource is represented in terms of Lemon, LexInfo and PREMON ontologies and will we published at Linguistic Linked Open Data cloud.

Keywords: FrameNet, Tatar language, Linguistic Linked Open Data.

1   Introduction

In this paper, we present TatVerbBank, the first FrameNet-like resource for Tatar language. This project is inspired by FrameNet and FrameBank [1].

Russian FrameBank is a bank of annotated samples with lexical constructions (e.g. argument constructions of verbs and nouns) from the Russian National Corpus. FrameBank belongs to FrameNet-oriented resources, but unlike Berkeley FrameNet it focuses more on morphosyntactic and semantic features of individual lexemes rather than on generalized frames.

In FrameNet the central element is the frame, but in FrameBank the lexeme is the central element and individual lexeme has its own set of lexical constructions. The resource directory contains a list of lexemes, but not frames. Each sense of a lexeme can be represented by a unique frame.

Information about each lexical construction in FrameBank is stored as a construction template, which includes [1]:

1. the syntactic rank of the element (Subject, Object, Predicate, Peripheral, Clause);
2. the morphosyntactic features of the element (including POS, case and preposition marking);
3. its status: lexical constant vs. variable;
4. the semantic roles of the argument (e.g., Agent, Patient, Instrument);
5. the lexical-semantic class of the element (e.g., human, animate, abstract entity, means of transport, etc.);

6. the morphosyntactic features of the target lexical unit itself (e.g. impersonal, passive participle, etc.);
7. one or several examples.

FrameBank serves as a main prototype for developing the Tatar VerbBank resource (hereinafter TatVerbBank).

The source of data for the VerbBank resource is the "Tugan tel" Tatar National Corpus (http://tugantel.turklang.tatar). Taking the descriptions of the model of constructions in FrameBank as the basis, we are taking into account the specific characteristics of the Tatar language. Our goal is to create a dictionary of verb constructions with semantic and especially syntactic information about verbal actants.

We develop our dictionary based on valency grammar as syntactic theory. At this stage, we use a reduced model of construction descriptions in Tatar compared with FrameBank, which includes:

1. the syntactic rank of the element (Subject, Object, Predicate, Peripheral, Clause);
2. the morphosyntactic features of the element (including POS, affix marking);
3. the semantic roles of the argument (e.g., Agent, Patient, Instrument);
4. the lexical-semantic class of the element (e.g., human, animate, abstract entity, means of transport, etc.);
5. one or several examples.

2 Semantic Roles in FrameBank and TatVerbBank

A base hierarchy of predicates and semantic roles is defined in FrameBank. The detailed list of semantic roles in this resource currently contains 91 items classified into seven domains such as Agent, Possessives, Patient, Addressee, Experiencer, Instrument and Settings. These domains are further subdivided into smaller units. It allows selecting different sets of semantic roles in individual lexical constructions.

For example, the role of Agent is defined as an active (prototypically animate) participant of a situation, intentionally changing something in the world. However there are more specific verbs with their own semantic roles (Speaker, Subject of motion, Subject of social relationship, etc.). These semantic roles are included in the hierarchy of agents and are linked with the predicates of the corresponding thematic classes. TatVerbBank uses the same set of semantic roles as defined in FrameBank. Core roles uniquely define frames, and peripheral roles are used to describe aspects of events in general.

3 TatVerbBank Description

When building TatVerbBank, we are using various lexical resources for the Tatar language. It should be noted that there is a lack of semantic dictionaries for the Tatar language. The main lexical resource is the Russian-Tatar explanatory dictionary by F.A.

Ganiev. Another lexical resource is the Russian-Tatar Social-Political thesaurus developed in Institute of Applied Semiotics of Tatarstan Academy of Sciences (http://tattez.turklang.tatar). At the first stage, we developed the verb dictionary which consists of Tatar lexemes denoting events, phenomena or processes. Then we grouped words into "sense groups" and built a proto-structure (proto-frame) for each group. The verbs (concepts) from the verb dictionary can be ambiguous and have different senses. Then we linked concepts of this dictionary with concepts of the Russian-Tatar Social-Political thesaurus to find ambiguous concepts.

The TatVerbBank resource unit is represented by a coherent structure with appropriate semantic and syntactic frames, as well as thesaurus concepts. The example of the TatVerbBank resource unit is shown on figure 1.

Here, the concept *GRANT* has two hypernyms (*PAYMENT* and *AID*) in the thesaurus. Also, for the given concept gives lexical units (text inputs) in Russian and Tatar (verb *субсидировать* (ru)/ analytical verb *субсидия бирү* (tat) /*to grant from the budget* (en)). The frame *GRANT* has core semantic roles as *PAYER, RECIPIENT* and *SUBSIDY*. Each core semantic role has its own case form as *PAYER* in nominative case, *RECIPIENT* in dative case, *SUBSIDY* in accusative case.
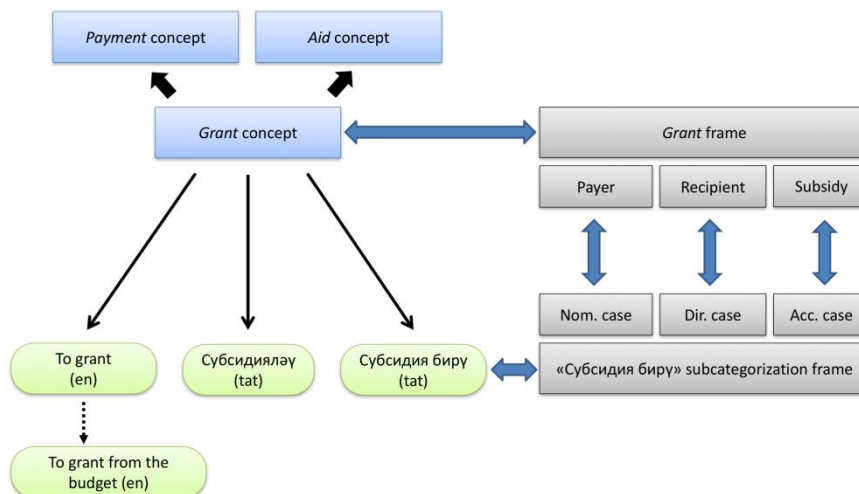

Fig. 1. Data Relationships in the TatVerbBank Model

4   TatVerbBank in Linguistic Linked Open Data Cloud

The resource is intended to be integrated into Linguistic Linked Open Data cloud [2] and is represented in terms of Lemon [3], LexInfo, OLiA [4] and PREMON [5] ontologies as well as a new custom ontology.

The lexical entries are represented as instances of ontolex:LexicalEntry class, syntactic frames as instances of synsem:SyntacticFrame class and semantic frames as instances of pmo:SemanticClass class. The frames are interlinked with RuThes Cloud [6] concepts and lexical entries are interlinked with TatThes [7] entries.

Fig. 2 depicts an example of оч 'to fly' lexical entry, the syntactic frame of this entry and its arguments as well as the Fly semantic frame and its frame elements, mapped to the syntactic arguments.

```
<http://lod.ruthes.org/resource/entry/TT-оч-v>
  a ontolex:LexicalEntry;
  rdfs:label "оч"@tt;
  lexinfo:partOfSpeech lexinfo:verb;
  ontolex:canonicalForm
    <http://lod.ruthes.org/resource/form/TT-очу-v-ger>;
  ontolex:evokes
    <http://lod.ruthes.org/resource/semantic-frame/fly-1>;
  synsem:synBehavior
    <http://lod.ruthes.org/resource/syntactic-frame/TT-оч-v-frame>.

<http://lod.ruthes.org/resource/form/TT-очу-v-ger>
  a ontolex:Form;
  ontolex:writtenRep "очу"@tt;
  lexinfo:verbFormMood lexinfo:gerunditive.

<http://lod.ruthes.org/resource/semantic-frame/fly-1>
  a pmo:SemanticClass;
  rdfs:label "Fly frame"@en;
  skos:broader <http://lod.ruthes.org/resource/semantic-frame/move-1>;
  pmo:semRole
    <http://lod.ruthes.org/resource/frame-element/fly-1@agent>,
    <http://lod.ruthes.org/resource/frame-element/fly-1@source>,
    <http://lod.ruthes.org/resource/frame-element/fly-1@goal>,
    <http://lod.ruthes.org/resource/frame-element/fly-1@place>,
    <http://lod.ruthes.org/resource/frame-element/fly-1@purpose>.

<http://lod.ruthes.org/resource/frame-element/fly-1@agent>
  a pmo:SemanticRole;
  tvbo:thematicRole tvbo:agent.

<http://lod.ruthes.org/resource/frame-element/fly-1@source>
  a pmo:SemanticRole;
  tvbo:thematicRole tvbo:source.

<http://lod.ruthes.org/resource/frame-element/fly-1@goal>
```

```
  a pmo:SemanticRole;
  tvbo:thematicRole tvbo:goal.

<http://lod.ruthes.org/resource/frame-element/fly-1@place>
  a pmo:SemanticRole;
  tvbo:thematicRole tvbo:place.

<http://lod.ruthes.org/resource/frame-element/fly-1@purpose>
  a pmo:SemanticRole;
  tvbo:thematicRole tvbo:purpose.

<http://lod.ruthes.org/resource/syntactic-frame/TT-оч-v-frame>
  a synsem:SyntacticFrame;
  synsem:synArg
    <http://lod.ruthes.org/resource/syntactic-argument/TT-оч-v-frame-arg-1>,
    <http://lod.ruthes.org/resource/syntactic-argument/TT-оч-v-frame-arg-2>,
    <http://lod.ruthes.org/resource/syntactic-argument/TT-оч-v-frame-arg-3>,
    <http://lod.ruthes.org/resource/syntactic-argument/TT-оч-v-frame-arg-4>,
    <http://lod.ruthes.org/resource/syntactic-argument/TT-оч-v-frame-arg-5>.

<http://lod.ruthes.org/resource/syntactic-argument/TT-оч-v-frame-arg-1>
  a synsem:SyntacticArgument;
  lexinfo:case lexinfo:nominativeCase;
  pmo:valueObj <http://lod.ruthes.org/resource/frame-element/fly-1@agent>.

<http://lod.ruthes.org/resource/syntactic-argument/TT-оч-v-frame-arg-2>
  a synsem:SyntacticArgument;
  lexinfo:case lexinfo:ablativeCase;
  pmo:valueObj <http://lod.ruthes.org/resource/frame-element/fly-1@source>.

<http://lod.ruthes.org/resource/syntactic-argument/TT-оч-v-frame-arg-3>
  a synsem:SyntacticArgument;
  lexinfo:case lexinfo:allativeCase;
  pmo:valueObj <http://lod.ruthes.org/resource/frame-element/fly-1@goal>.

<http://lod.ruthes.org/resource/syntactic-argument/TT-оч-v-frame-arg-4>
  a synsem:SyntacticArgument;
  lexinfo:case lexinfo:locativeCase;
  pmo:valueObj <http://lod.ruthes.org/resource/frame-element/fly-1@place>.

<http://lod.ruthes.org/resource/syntactic-argument/TT-оч-v-frame-arg-5>
  a synsem:SyntacticArgument;
  lexinfo:partOfSpeech lexinfo:verb;
  lexinfo:verbFormMood lexinfo:infinitive;
  tvbo:preposition <http://lod.ruthes.org/resource/entry/TT-диеп-prep>;
```

```
pmo:valueObj <http://lod.ruthes.org/resource/frame-element/fly-1@purpose>.
```

Fig. 2. The *очу* lexical entry and its syntactic and semantic frames

5   Conclusion

In this paper we presented TatFrameBank, the first FrameNet-like resource for Tatar
language.

The resource is under development now and our immediate goal is to release its
public version consisting of approximately 100 key verbs. After that, we are going to:
1. complement the frames by their realizations from the national corpus of Tatar
    language "Tugan Tel";
2. develop frames for less frequent verbs;
3. develop frames for other parts of speech and idiomatic phrases.

References

1. Lyashevskaya, O., Kashkin, E.: FrameBank: A Database of Russian Lexical Constructions.
    In: Khachay, M., et al (eds). Proceedings of the 4th International Conference on Analysis of
    Images, Social Networks and Texts (AIST 2015). Communications in Computer and Infor-
    mation Science, vol 542, pp. 350-360. Springer (2015). doi:10.1007/978-3-319-26123-2_34
2. Cimiano, P., Chiarcos, C., McCrae, J.P., and Gracia, J.: Linguistic Linked Open Data Cloud.
    In: Cimiano, P., et al. (eds.) Linguistic Linked Data: Representation, Generation and Appli-
    cations, pp. 29–41. Springer (2020). https://doi.org/10.1007/978-3-030-30225-2_3.
3. McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P.: The OntoLex-Lemon
    Model: Development and Applications. In: Kosem I., et al. (eds.) Proceedings of the 5th
    biennial conference on Electronic Lexicography (eLex 2017), pp. 587–597. Lexical Com-
    puting CZ (2017).
4. Chiarcos, C.: OLiA – Ontologies of Linguistic Annotation. Semantic Web **6**(4), 379–386
    (2015). https://doi.org/10.3233/SW-140167.
5. Rospocher, M., Corcoglioniti, F., and Palmero Aprosio, A.: PreMOn: LODifing linguistic
    predicate models. Language Resources and Evaluation **53**, 499–524 (2019).
    https://doi.org/10.1007/s10579-018-9437-8.
6. Kirillovich, A., Nevzorova, O., Gimadiev, E., Loukachevitch, N.: RuThes Cloud: towards a
    multilevel linguistic linked open data resource for Russian. In: Różewski, P., Lange, C.
    (eds.) KESW 2017. CCIS, vol. 786, pp. 38–52. Springer, Cham (2017).
    https://doi.org/10.1007/978-3-319-69548-8 4.
7. Galieva, A., Kirillovich, A., Khakimov, B., Loukachevitch, N., Nevzorova, O., Suleymanov,
    D.: Toward domain-specific Russian-tatar thesaurus construction. In: Proceedings of the In-
    ternational   Conference   IMS-2017,   pp.   120–124.   ACM   (2017).
    https://doi.org/10.1145/3143699.3143716.