Russian-Tatar Socio-Political Thesaurus: Methodology, Challenges, Status of the Project

Alfiya Galieva, Olga Nevzorova, Dilyara Yakubova

Tatarstan Academy of Sciences, Kazan, Russia

amgalieva@gmail.com

Turklang - 2017

Outline

- Introduction
- Methodology
- Challenges
- Synonymy in Tatar socio-political vocabulary
- Status of the project
- Conclusion

Introduction

For low-resource languages like Tatar, terminology compilation may be a challenge in many respects:

- because of the insufficient degree of terminology development that leads to absent relevant designations for special concepts in a language;
- because of lacking special sources for assembling terminology;
- because of lacking coherent terminology management at the terminology development stage as well as at the stage of usage, etc.

Introduction

Aim:

to compile a modern Tatar Thesaurus related to the following domains:
- *state government,*
- *economy,*
- *social   life,*
- *justice,*
- *warfare,*
- *culture,*
- *religion.*

Methodology

- The Russian-Tatar Thesaurus is being developed in the format of the RuThes thesaurus as a "network of concepts and their lexical entries with basic semantic relations mapped between them".

    http://www.labinform.ru/pub/ruthes/index.htm

- The conceptual structure of RuThes is mainly preserved,
- The Tatar component is based on the list of concepts of RuThes

Stages:

1. Search for equivalents that are actually used in Tatar as translations of Russian terms.
2. Adding new concepts.
3. Revising the relations between concepts.

Methodology

- During the first stage, RuThes concept names and text entries are translated into Tatar.
- The sources of translation: bilingual general-purpose dictionaries, special socio-political dictionaries, Tatar texts on socio-political topics, and available Russian-Tatar parallel texts.

Requirements to Tatar concept names:

- larity of meaning;
- use of neutral and common lexical units for concept names;
- brevitcy of denomination;
- selecting the most precise, the most frequent and officially used denomination from several options

Methodology

- Russian concept name and its Tatar equivalent may differ in length and and structure.
- Specifications for ambiguous words (meaning). Tatar specifications may differ from the Russian.

CONCEPT NAME IN RUSSIAN	CONCEPT NAME IN TATAR
РУКОВОДСТВО (УЧЕБНОЕ ПОСОБИЕ) 'guidance, manual'	КУЛЛАНМАЛЫК 'guidance, manual'
СТОИМОСТЬ 'cost'	КЫЙММӘТ (ЧЫГЫМНАР КҮЛӘМЕ) 'cost (an amount of expenses)'

Methodology

When compiling the list of Tatar terms (lexical entries), we use data of available corpora built for the Tatar language:

1. The Corpus of Written Tatar (CWT) by researchers of Kazan Federal University;

2. "Tugan Tel" Tatar National Corpus (TNC) by the Institute of Applied Semiotics of Tatarstan Academy of Sciences - http://tugantel.tatar/?lang=en

- Socio-Political Subcorpus - http://tugantel.tatar/corpus/op/
-

Challenges

Stage 1. Search for equivalents that are actually used in Tatar:
- the dictionaries are incomplete and do not contain entries reflecting present-day socio-political items widely used in Russia (such as *political correctness, framework agreement, taxable*…)
- the dictionaries do not contain Tatar translation units in actual use, but those in use in the Soviet era.

Use of media texts, texts from official documents

Challenges

Stage 2. Adding new concepts related to issues of political, religious and cultural life of the Tatar society

Stage 3. Revising relations among the concepts
Concepts added at a higher level:
- OFFICIAL HOLIDAY OF TATARSTAN REPUBLIC
- MUSLIM HOLIDAY

Concepts added at a lowerer level to CALENDAR DATE:
- HIJRI CALENDAR DATE
- GREGORIAN CALENDAR DATE

Challenges

Example: **УРАЗА ГАЕТЕ** ('Eid Al-Fitr')

CONCEPTS ABOVE	ASSOCIATED CONCEPTS	TEXT ENTRIES
ДИНИ БӘЙРӘМ 'religious holiday'	Association 1: УРАЗА 'Muslim fasting'	ураза гаете, ураза бәйрәме, рамазан гаете, рамазан бәйрәме
ТАТАРСТАН РЕСПУБЛИКАСЫНЫҢ РӘСМИ БӘЙРӘМЕ 'official holiday of the Republic of Tatarstan'	Association 2: РАМАЗАН 'Ramadan'	
МӨСЕЛМАН ДИНЕ БӘЙРӘМЕ 'Muslim Holiday'	Association 3: УРАЗА ФИТЫРЫ 'Za-kat al-Fitr'	
ҺИҗРИ КАЛЕНДАРЬ ДАТАСЫ 'Hijri calendar date'		
КАЛЕНДАРЬ ДАТАСЫ 'calendar date'		

SYNONYMY

Abundance of synonyms

1. Due to the **borrowings** from:
- the Arab-Muslim cultural area
- the European cultural area

Words of different origin for the same referent (Turkic and Tatar, Arabic and Persian, Greek, Latin, English and Russian)

Diverse groups of specialists with dissimilar world-views and ideological guidelines develop Tatar terminology

Distribution of synonyms coming from different languages

Lexeme	English translation, PoS	Origin of the word	Number in the CWT	Number in the TNC
экономик	Economic	Greek	621	676
икътисади		Arabic	27,351	22,274
политик	Political (adj.)	Greek	1,008	1,535
сәяси		Arabic	25,011	24,489
суд	Court of law (noun)	Russian	15,725	14,243
мәхкәмә		Arabic	8,718	8,101

SYNONYMY

Competing synonyms of Tatar origin

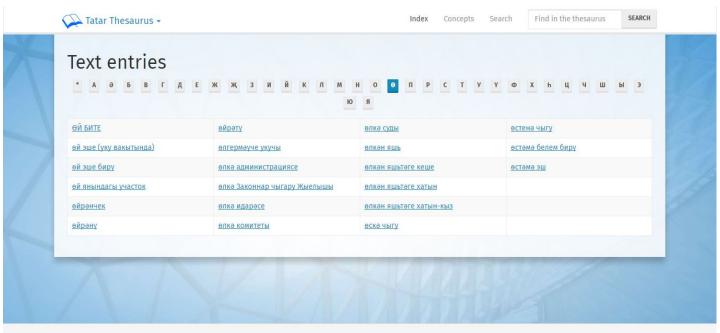| Lexeme	English translation	TNC
эшкуар	businessman	11,280
эшмәкәр		13,876
кәсәбәче		112

SYNONYMY

Abundance of synonyms

2. Due to **peculiarities of lexical, derivational and grammatical systems** of the Tatar language

Example of regularly corresponding grammatical patterns of noun phrases: *ADJ +N* and *N + N, POSS_3*

икътисадый кризис (ADJ +N),
икътисад кризисы (N + N, POSS_3)
'economic crisis'.

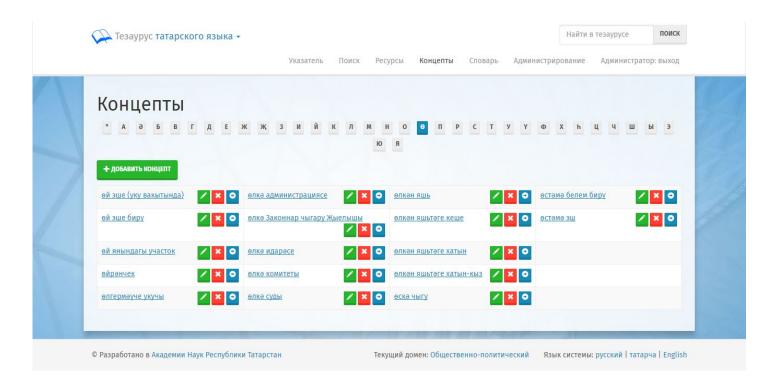- 9 noun phrases to designate the term JOINT-STOCK COMPANY found in the media and official texts

Status of the Project

- **6,500 concepts**
- Web application: http://tattez.antat.ru/

Open part

Closed part

Conclusion

- Building a bilingual Russian-Tatar Thesaurus is an important step for Tatar terminology management

- Future:
- expand the thesaurus;
- methods of extracting domain-specific terms from texts and corpus data developing.

Acknowledgements

Thank you very much for your attention!