9th International Conference on Computer Processing of Turkic Languages
TurkLang 2021, Kyzyl, Tuva, Russia, September 21-23, 2021

Cognitive studies of the lexico-grammatical potential of the Tatar language to create new information processing technologies

Dzhavdet Suleymanov

Kazan Federal University,
Institute of Applied Semiotics of TAS ips.antat.ru
Kazan, Russia, e-mail: dvdt.slt@gmail.com

INTRODUCTION: prehistory

- 1985 Tatar localization of the first personal computers, development of computer terminology
- 1990 Artificial Intelligence Laboratory

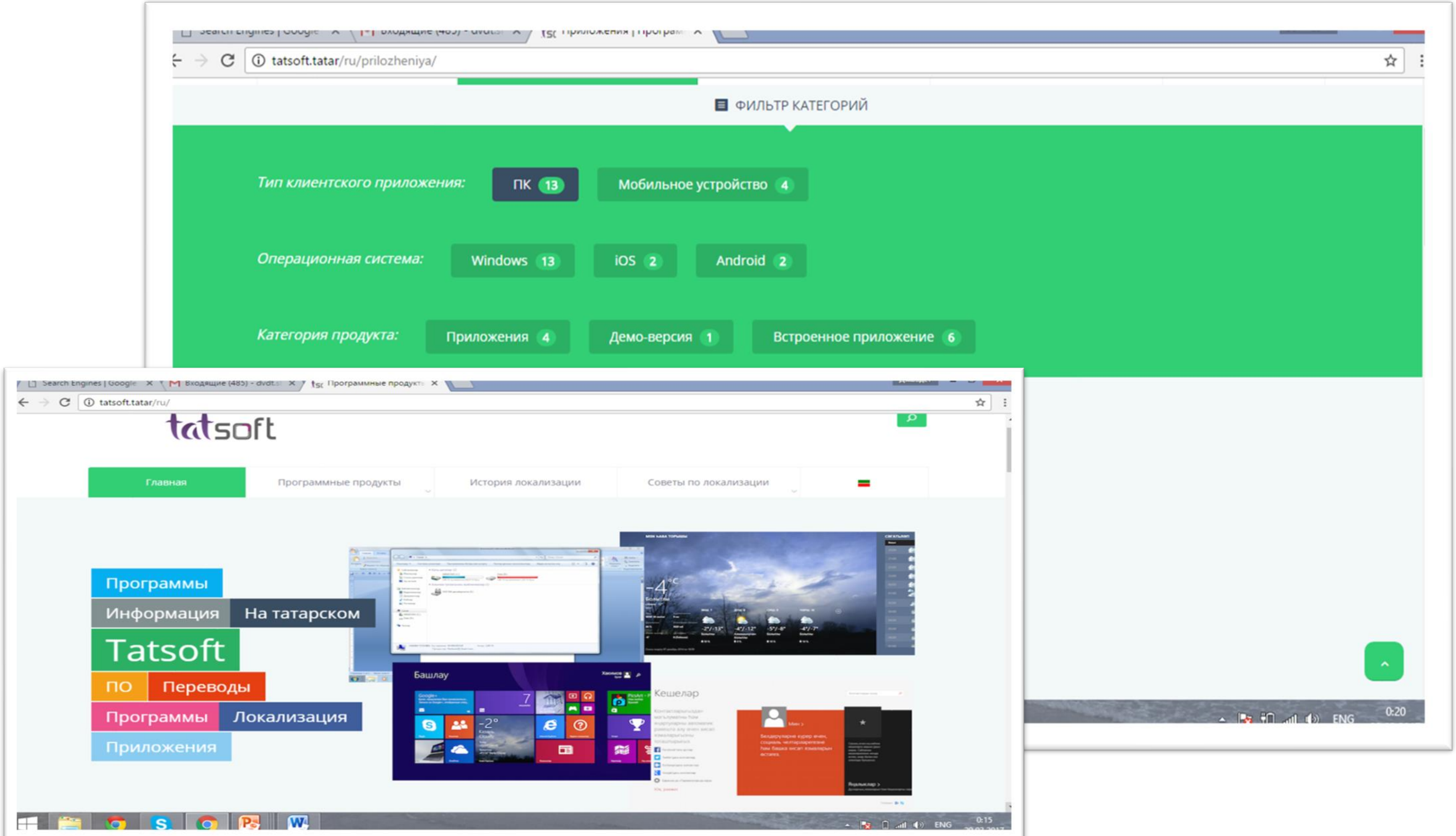
  The beginning of computer and mathematical linguistics at the Kazan University.

  The beginning of the development of the Computer Fund of the Tatar language.
- 1993 Joint Research Laboratory of AI at the Department of Theoretical Cybernetics: Tatar PC localization; development of applied programs; development of linguistic resources.
- 2009 Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan. Fundamental and applied research in the field of semiotic modeling, computational and cognitive linguistics.

INTRODUCTION: Research and Development

- National localization of the information and communication technology (TL in ICT)
- The development and adaptation of Information Technology for the TL (Software tools and Linguistic resources) (IT for TL)
- Research of the potential of the TL as a source for development of the new information technology (TL for ICT)

Software products and linguistic resources (https://tatsoft.tatar)



tatsoft.tatar/ru/prilozheniya/

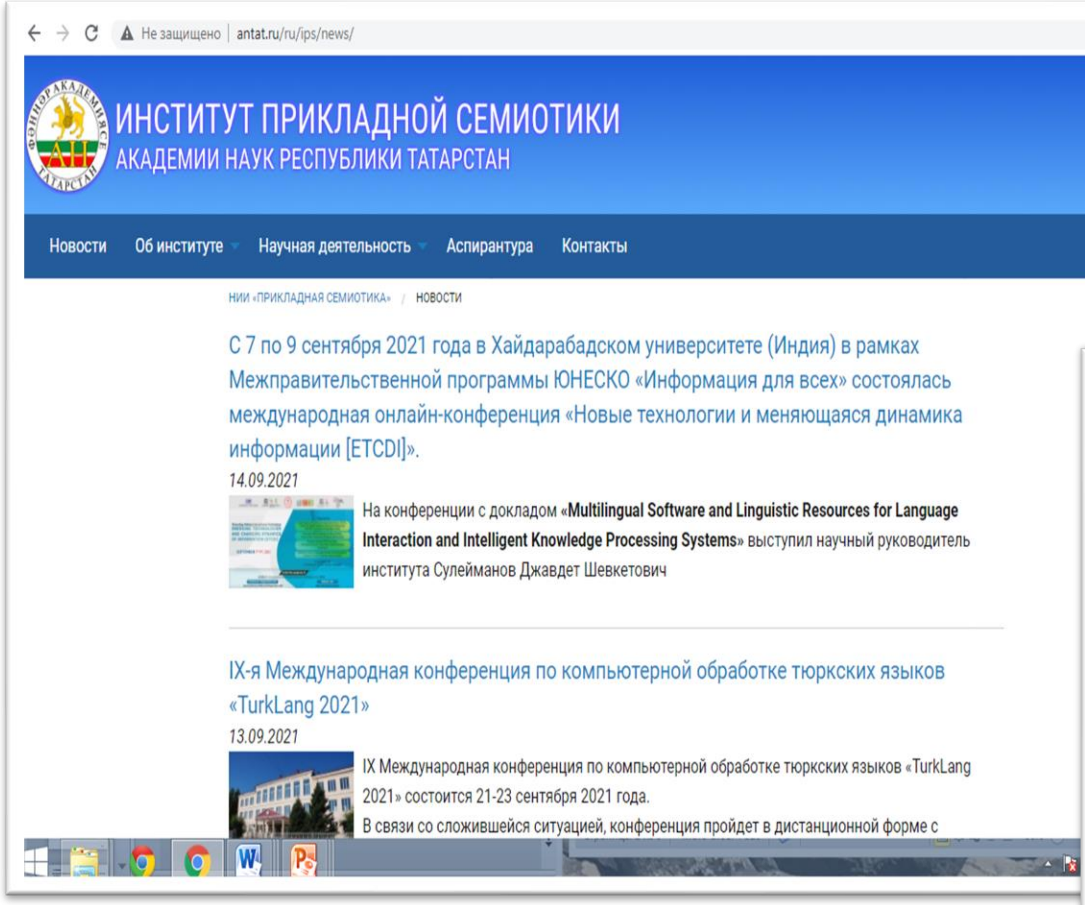
ФИЛЬТР КАТЕГОРИЙ

Тип клиентского приложения: ПК 13 | Мобильное устройство 4

Операционная система: Windows 13 | iOS 2 | Android 2

Категория продукта: Приложения 4 | Демо-версия 1 | Встроенное приложение 6

The Institute of Applied Semiotics website: ips.antat.ru


Не защищено | antat.ru/ru/ips/news/
РУС TAT ENG

ИНСТИТУТ ПРИКЛАДНОЙ СЕМИОТИКИ
АКАДЕМИИ НАУК РЕСПУБЛИКИ ТАТАРСТАН

Новости
Об институте
Научная деятельность
Аспирантура
Контакты
НИИ «ПРИКЛАДНАЯ СЕМИОТИКА» / НОВОСТИ

С 7 по 9 сентября 2021 года в Хайдарабадском университете (Индия) в рамках Межправительственной программы ЮНЕСКО «Информация для всех» состоялась международная онлайн-конференция «Новые технологии и меняющаяся динамика информации [ETCDI]».

14.09.2021

На конференции с докладом «**Multilingual Software and Linguistic Resources for Language Interaction and Intelligent Knowledge Processing Systems**» выступил научный руководитель института Сулейманов Джавдет Шевкетович

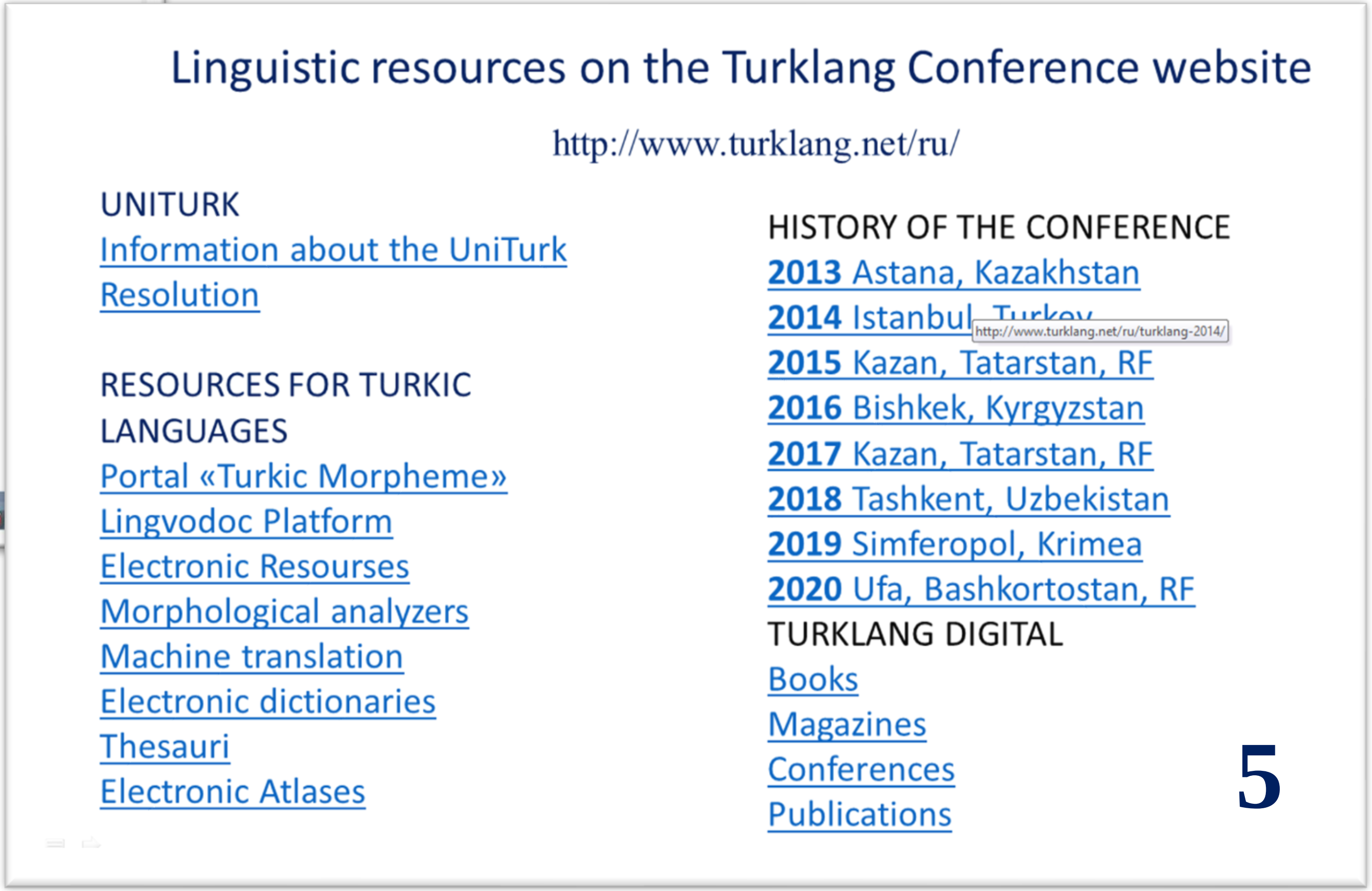
IX-я Международная конференция по компьютерной обработке тюркских языков «TurkLang 2021»

13.09.2021

IX Международная конференция по компьютерной обработке тюркских языков «TurkLang 2021» состоится 21-23 сентября 2021 года.

В связи со сложившейся ситуацией, конференция пройдет в дистанционной форме с

Turklang Conference website: www.turklang.net/ru



Linguistic resources on the Turklang Conference website

http://www.turklang.net/ru/

UNITURK
Information about the UniTurk
Resolution

RESOURCES FOR TURKIC LANGUAGES
Portal «Turkic Morpheme»
Lingvodoc Platform
Electronic Resourses
Morphological analyzers
Machine translation
Electronic dictionaries
Thesauri
Electronic Atlases

HISTORY OF THE CONFERENCE
2013 Astana, Kazakhstan
2014 Istanbul, Turkey
2015 Kazan, Tatarstan, RF
2016 Bishkek, Kyrgyzstan
2017 Kazan, Tatarstan, RF
2018 Tashkent, Uzbekistan
2019 Simferopol, Krimea
2020 Ufa, Bashkortostan, RF

TURKLANG DIGITAL
Books
Magazines
Conferences
Publications

Third Direction of Research and Development - Tatar for ICT

- Development of intelligent operating systems, programming languages and intelligent software tools
- Development of communication language with artificial intelligence systems using structural, conceptual and cognitive characteristics of the Tatar language as a language of agglutinative type
- The solution to this task will ensure the use of the Tatar language in cyberspace on the new, on a qualitatively higher, motivational level. Obviously, it will increase interest in studying the potential of the language from the position of creating new technological capabilities for the purposes of the cyberspace itself.

THREE ASPECTS OF NATURAL LANGUAGES INVESTIGATION

- The cognitive aspect – facilities of NL for describing the “world model” and for representing knowledge.

- The communicative aspect – facilities of NL for encoding, receiving and sending, processing the information and for supporting the dialogue.

- The technological aspect - facilities of NL for implementing the means for effective processing, adequate describing and compact storing of information, developing ergonomic technical means for developing intellectual program tools.

Interest in the potential of natural languages

Artificial programming languages are based on deep cognitive structures of NL, and hence their mentality. Therefore, these systems implement the descriptive and computational potential of the corresponding NL.

It is of great interest to study the lexico-grammatical features (morphological, syntactical, semantic) and to determine technological effectiveness of Tatar Language in order to develop software tools for effective processing of NL information.

Areas of research for development of new technologies

- research and identification of natural grammatical structures of the Tatar language, with almost regular grammar and natural complexity, in order to create a new generation of artificial intelligence languages on their basis;
- development of an intermediary language based on subsets and constructions of languages with certain cognitive properties, allowing the most adequate and concise description of the context and fast processing of texts in NL;
- the creation of cognitive models of TL, sufficiently relevant to reflect the mentality of the language, based on "common sense", which could be used as a basis for the creation of so-called explanatory artificial intelligence.

TECHNOLOGICAL FEATURES ARE MOST IMPORTANT FOR KNOWLEDGE PROCESSING SYSTEM AND TECHNOLOGY:

1. Time (minimization)
2. Memory (minimization)
3. Compactness of information storing and delivering
4. Possibilities for encoding and processing of fuzzy information
5. Knowledge activeness

(1-3) – determine effectiveness

(4-5) – determine intellectuality of systems and technologies

LEXICAL AND GRAMMATICAL FEATURES OF THE TATAR LANGUAGE

1. Morphological regularity
2. Agglutinativity (possibility to change the word form by gluing definite affixes)
3. Morphological ellipsis and recursion
4. Morphological (synthetic) means of expressing modality
5. Contextual variety of meanings of the affixes
6. Fuzziness of command and action description and describing a role situation with one verb word form
7. Knowledge activity

REGULARITY AND NATUARAL COMPLEXITY OF MORPHOLOGY

Characteristics of Tatar morphology: regularity

Formal grammars have minimal characteristics of time and memory functions for information processing.

Tatar language morphology minimizes memory and time functions in information processing.

REGULARITY OF MORPHOLOGY OF NOMINAL AND VERB FORMS

- (urman (forest), apa (sister), kosh (bird), ..., chəchək (flower)) + **LAR-Ybyz-GA-DYR-MY**:

urman**narybyzgadyrmy** - are in our forests?

apa**larybyzgadyrmy** - do our sisters have?

кosh**larybyzgadyrmy -** what is in our birds?

chəchə**klərebeezgəderme -** what is in our colors?

- (uyna (play), eshlə (work), tor (stop), ..., sal (remove)) + **dY-lAr-mY:**

uina**dylarmy** - did they play?

eshlə**delərme** - did the work?

tor**dylarmy** - did the stand?

sal**dylarmy -** were they removed?

AGGLUTINATIVENESS

FORMATION OF NEW WORD FORMS BY GLUING CERTAIN AFFIXAL MORPHEMES

Example:

Tatarchalashtyrgalashtyruchylardagynykylargamyni?

Tatar/cha/la/shtyr/gala/shtyr/u/chy/lar/dagy/nyky/lar/ga/myni?

("Is it to those who belong to what is on those who are involved in Tatar localization from time to time")

This word form has the folowing structure:

Tatar (noun) + ***cha*** (adverb) + ***la*** (verb) + ***shtyr*** (verb, mode) + ***gala*** (verb, mode) + ***shtyr*** (verb, mode)+***u*** (substantive, verb-name)+ ***chy*** (noun) + ***lar*** (plural) + ***dagy*** (substantive, locative) + ***nyky*** (substantive, possessive) + ***lar*** (plural) + ***ga*** (directive) + ***myni*** (question, surprise).

MORPHOLOGICAL ELLIPSIS

Example:

- *Min kyr**larybyzga**, urman**narybyzga**, jylga**larybyzga**, tau**larybyzga** shatlanam = Min kyr, urman, jylga, tau**lar-y-byz-ga** shatlanam. ("I rejoice at our camps, forests, rivers, mountains".)*

RECURSION

Possibility of cyclic new meaning creation by consecutive application of the same formula.

Example 1: Lexeme **"tau"** ('mountain') **+ *–dagy*** = new <u>indefinite</u> objects or features:

- *<u>taudagy</u> – 'something on the mountain'; <u>taudagydagy</u> – 'something on something on the mountain';*
- *<u>taunyky</u> – 'something which belongs to the mountain'; <u>taunykynyky</u> – 'something which belongs to that which belongs to the mountain".*

Example 2:***Taunykyndagynykyndagynykyndagy***
Tau/nyky/ndagy/nyky/ndagy/nyky/ndagy
'tau' (noun+possessive+locative 2+possessive+ locative 2+ possessive +locative 2)
This word form means:

"something which is situated on (in) something which belongs to something which is situated on (in) something which belongs to something which is situated on (in) something which belongs to the mountain"

RECURSION

Example 3:

Every indefinite affix is followed by parameters:

tau+dagy(x1)+ndagy(x2)+nyky(x3)+nyky(x4)+ndagy(x5)+nyky(x6),
where *xi* – contextual objects (*i=1,6*).

Assigning the meanings to parameters:

X0=“tau” (‘mountain’), *x1=“peschera”* (“cave”), *x2=“medved”* (“bear”), *x3=“lapa”* (“paw”),

x4=“kogot” (“claw”), *x5=“myod”* (“honey”), *x6=“vkus”* (“taste”),

… we get the following contextual meaning:

“the taste which belongs to the honey which is situated on the claw on the paw which belongs to the bear which is situated in the cave wich is situated in the mountain” (“the taste of honey on a bear’s paw in the cave situated in the mountains”).

FUZZINESS OF COMMANDS AND ACTIONS

Example:

- ***u*** (“wash”) – “to wash” (person 3, singular, imperative);
- ***ugala*** (“wash from time to time”) – *u* (“wash”)+*gala* (“from time to time”);
- ***ugalashtyr*** (“wash from time to time, from time to time: more seldom”) – *u* (“wash”)*+ gala* (“from time to time”)*+ shtyr* (“from time to time”);
- ***ugalashtyrgala** (*“wash from time to time, from time to time, from time to time: even more seldom”*)* –

  u (“wash”)*+ gala* (“from time to time”)*+ shtyr* (“from time to time”) + *gala* (“from time to time”); **etc.**

KNOWLEDGE ACTIVITY

English sentences are formed according to the scheme:

(1) **S-V-O** (subject-verb-object)

Tatar sentences are formed according to the scheme:

(2) **S-O-V** (subject-object-verb)

Scheme (1): action controls the situation.

Scheme (2): action takes part after situation analysis.

KNOWLEDGE ACTIVITY

In the Tatar language first information is given and only then the action itself is defined

– whether positive or negative.

Example: *Min dustim belen irtege toshten song bulasi “Atilla” kinosina baram/barmiym* - lit: “I and my friend tomorrow after dinner to the film “Atilla” will (not) go”.

English: *My friend and I will not go to the movie «Attila», which will take place tomorrow afternoon.*

KNOWLEDGE ACTIVITY

- Natural style of thinking for intellectual systems is:

analysis-action, data-algorythms

- The command style is typical of modern technologies based on the English language mentality:

action-analysis, algorythm-data

Knowledge activity is natural for the Tatar language and is determined by its grammar.

Therefore, intellectual programmes on knowledge accumulating and extracting can be naturally based on the Tatar language grammar.

CONCLUSION

- At present, on the basis of our research, the project «Structure of a pragmatically-oriented model of natural language of agglutinative type has been developed based on the study of its cognitive aspects (using the example of the Tatar language)» and applied for a grant from the Russian Science Foundation.

- Scientific problem to be solved by the project: Research and Development of mathematical models that are relevant to the lexical and grammatical structure of the Tatar language, explicitly and adequately reflecting the potential of the language in the communicative and cognitive aspects.

CONCLUSION

- The aim of the project is to develop the structure of mathematical models that are relevant to the lexical and grammatical structure of the NL of the agglutinative type, explicitly and adequately reflecting the potential of the language in the communicative and cognitive aspects.
- It is planned to use the results obtained in the following promising areas:

1. Creation of a unified environment for research and processing of materials on agglutinative languages as the most suitable for the application of computer and logical methods.
2. Development and filling of a generalized model of the Tatar language as an example of the possibility of creating intelligent systems for processing natural languages on the basis of a focal-decentralized approach
3. Creation of a prototype of a universal language for the exchange of information between Artificial Intelligence systems, as well as between them and humans.

THANK YOU!

СПАСИБО!

БАЙЫРЛЫГ!

РӘХМӘТ!