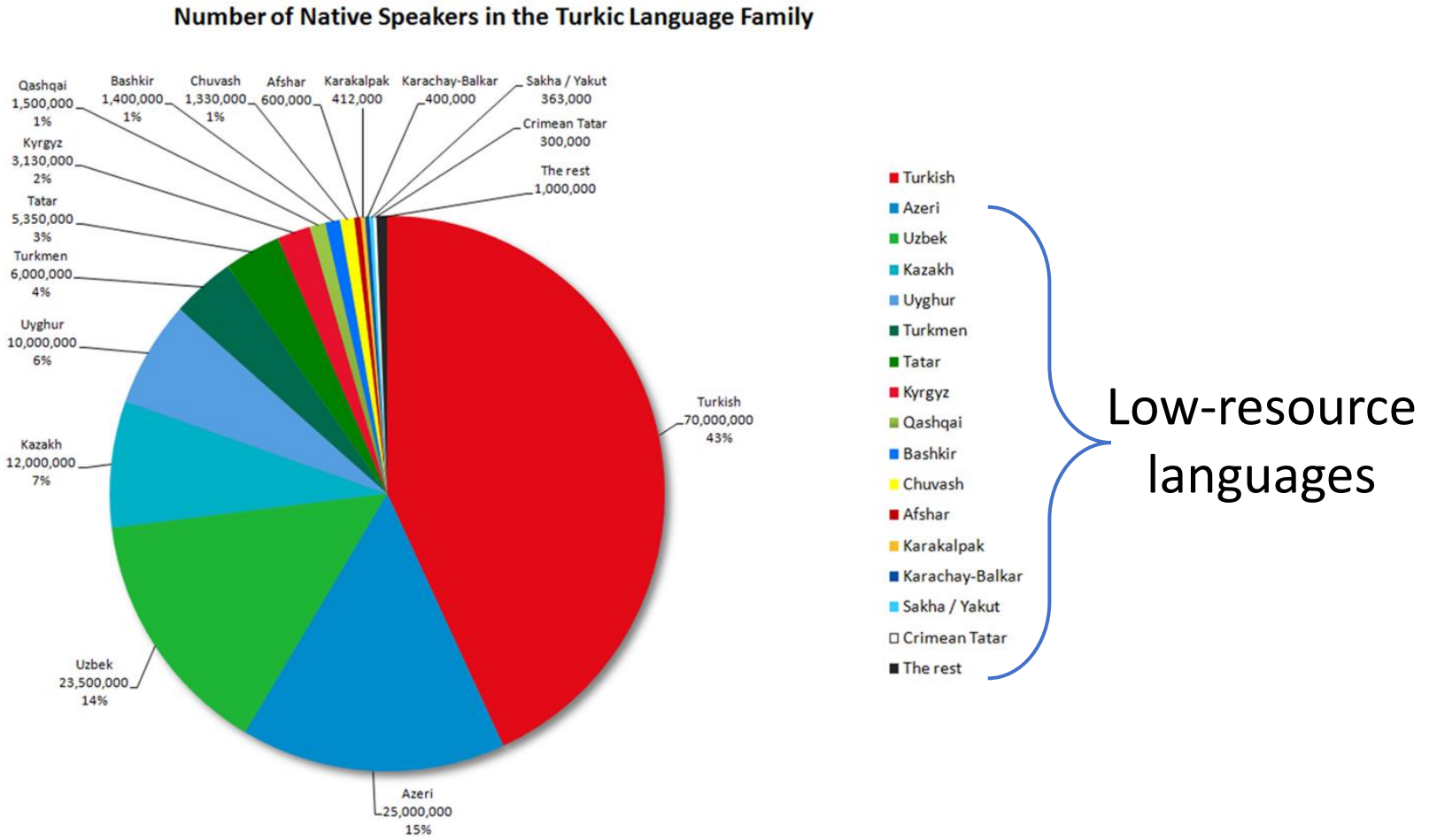
IX International Conference on Computer Processing of Turkic Languages “TurkLang 2021”
September 21-23, 2021, Kyzyl, Tuva

Internet portal "Turkic Morpheme" for the creation of multifunctional computer funds of closely related languages

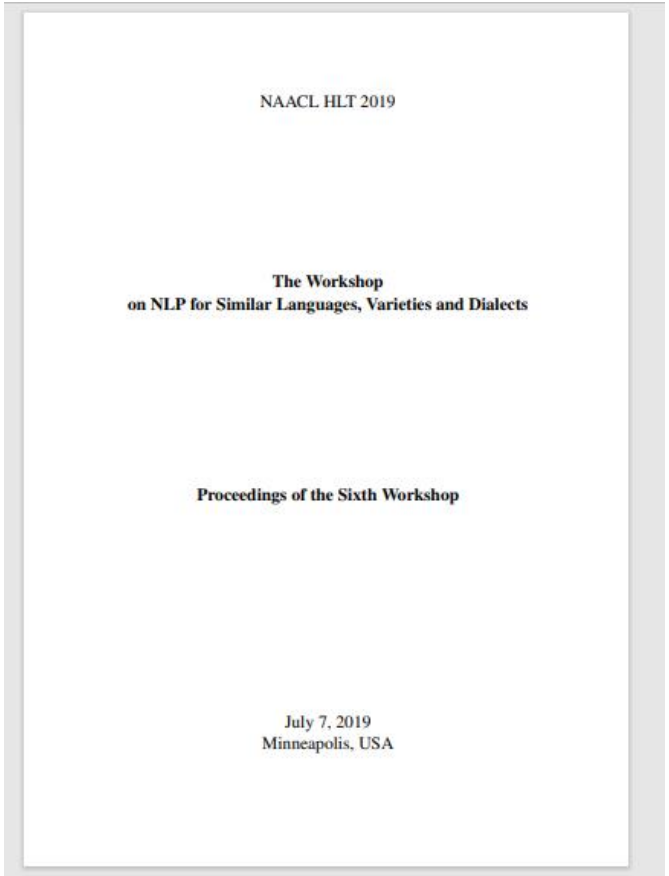
Ayrat Gatiatullin, Dzhavdet Suleymanov, Nikolai Prokopyev,
Nilufar Abdurakhmonova

Institute of Applied Semiotics
Academy of Sciences of the Republic of Tatarstan
Kazan Federal University
Russian Federation
Mirzo Ulugbek National University Of Uzbekistan

Actual problem



Solutions of the problem



NAACL HLT 2019

The Workshop
on NLP for Similar Languages, Varieties and Dialects

Proceedings of the Sixth Workshop

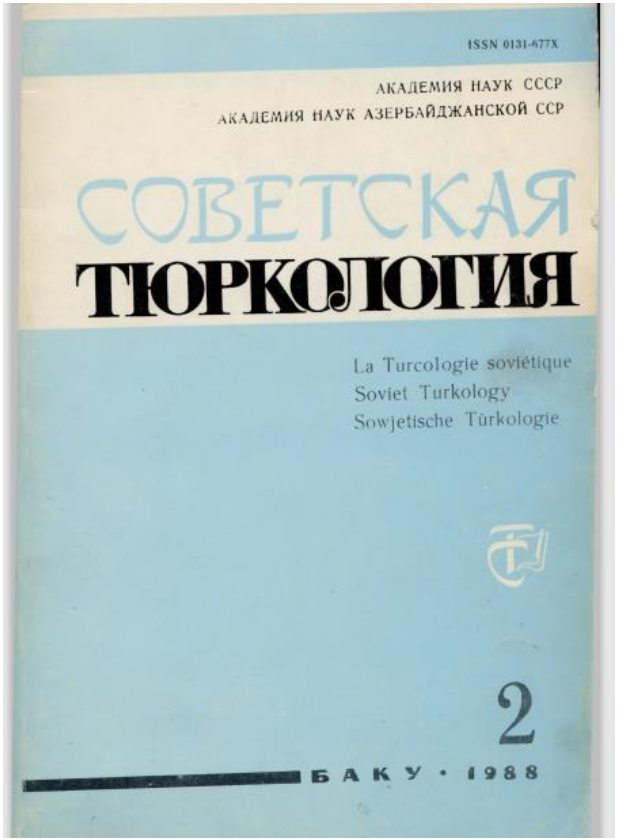
July 7, 2019
Minneapolis, USA

EMNLP' Workshop on Language Technology for Closely Related Languages and Language Variants

Preface:

Developing language resources and tools from scratch is expensive. Nevertheless, such efforts can often be reduced by using existing resources and tools for structurally and lexically interconnected and more resource-intensive languages.

History of the problem



V.G. Guzev, R.G. Piotrovsky, A.M. Shcherbak On the creation of the machine fund of the Turkic languages // Soviet Turkology. 1988. No. 2. S.92-101.

To solve practical problems there is a need to create a Large Multipurpose Computer Fund for Turkic Languages (LMC FTL), which should be built in a way of modelling both common Turkic language system and each specific language system (functioning or extinct, contemporary or ancient) with all inventory and structure elements, rules of sign representation of language elements in speech, including rules of linear interconnection of speech elements.

Discussion on the problem

TurkLang

International Conference on Computer Processing of Turkic Languages "TurkLang"

Home | News | Organizers | Submissions | Search | Find | EN

The IX International Conference on Computer Processing of Turkic Languages "TurkLang 2021"

Dear Colleagues,

We are glad to inform you that the VIII International Conference on Computer Processing of Turkic Languages "TurkLang 2021" will be held on September 21-23, 2021. Due to the current situation, the conference will be held remotely using modern information technologies.

Paper template for conference proceedings (RSCI) papertemplate_TurkLang2021_en.doc

Paper template for CMLS workshop proceedings (Scopus) spinproc1703.zip

Papers are to be uploaded in EasyChair: https://easychair.org/conferences/?conf=turklang2021

Conference will be held using ZOOM.
URL to connect: https://us02web.zoom.us/j/81314007818?pwd=bndjOUpXcXNxSUFFR3d2LzVEeGUvdz09
Conference ID: 813 1400 7818
Access code: 935601

Conference program: Program-TurkLang-2021-En.xlsx

CONFERENCE HISTORY

2013	Astana, Kazakhstan
2014	Istanbul, Turkey
2015	Kazan, Tatarstan, RF
2016	Bishkek, Kyrgyzstan
2017	Kazan, Tatarstan, RF
2018	Tashkent, Uzbekistan
2019	Simferopol, Crimea, RF
2020	Ufa, Bashkortostan, RF

UNITURK

- Information about UniTurk
- Resolution of UniTurk

RESOURCES FOR TURKIC LANGUAGES

- Portal "Turkic Morpheme"
- Platform Lingvodoc
- Electronic Corpora
- Morphological analyzers
- Systems of Machine Translation
- Electronic dictionaries
- Thesaurus
- Electronic atlases

TURKLANG DIGITAL

- Books
  - Books about Ortaturk
- Magazines
- Conferences
- Publications

Information about resources

Electronic Corporas

Corpus of Altay language	http://altay2.gasu.ru/
Corpus of Bashkir language (poetic)	http://web-corpora.net/bashcorpus/search/
Corpus of Kazakh language	http://kazcorpus.kz
Corpus of Crimeantatar language	http://korpus.juls.savba.sk/QIRIM/
Corpus of wtitten Tatar	http://www.corpus.tatar/
Corpus of Tatar language 'Tugan tel'	http://tugantel.tatar/
Corpus of Tuvinian language	http://www.tuvancorpus.ru/
Corpus of Turkish language	http://www.tnc.org.tr/
Corpus of Uzbek language	http://uzbekcorpus.uz/enVer
Corpus of Khakas language	http://khakas.altaica.ru/
Corpus of Shor language	http://corpora.iea.ras.ru/corpora/

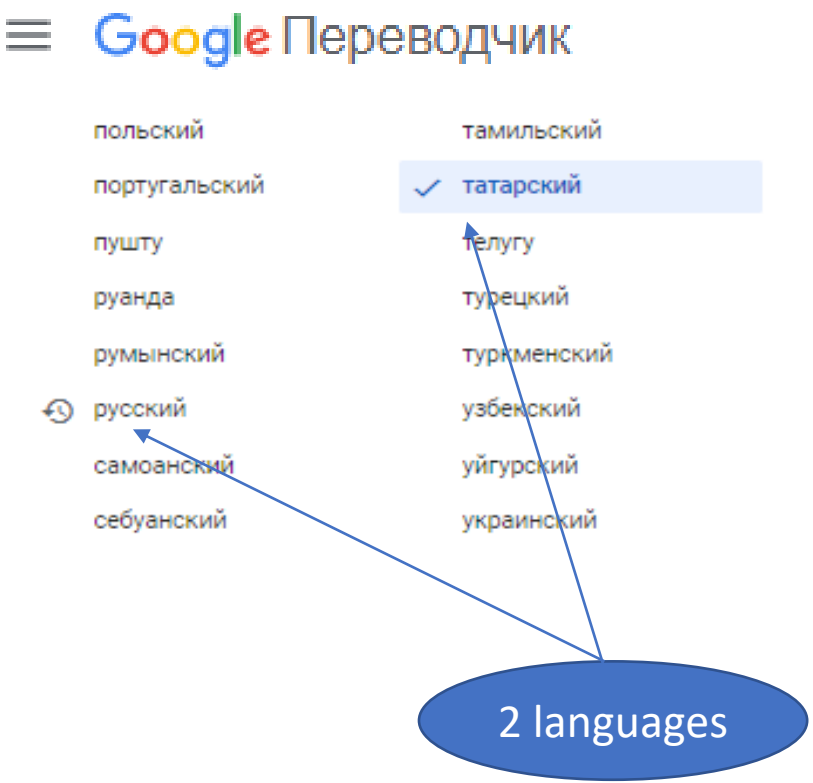
CONFERENCE HISTORY

2013	Astana, Kazakhstan
2014	Istanbul, Turkey
2015	Kazan, Tatarstan, RF
2016	Bishkek, Kyrgyzstan
2017	Kazan, Tatarstan, RF
2018	Tashkent, Uzbekistan
2019	Simferopol, Crimea, RF
2020	Ufa, Bashkortostan, RF

UNITURK

- Information about UniTurk
- Resolution of UniTurk

Commercial solutions to the problem

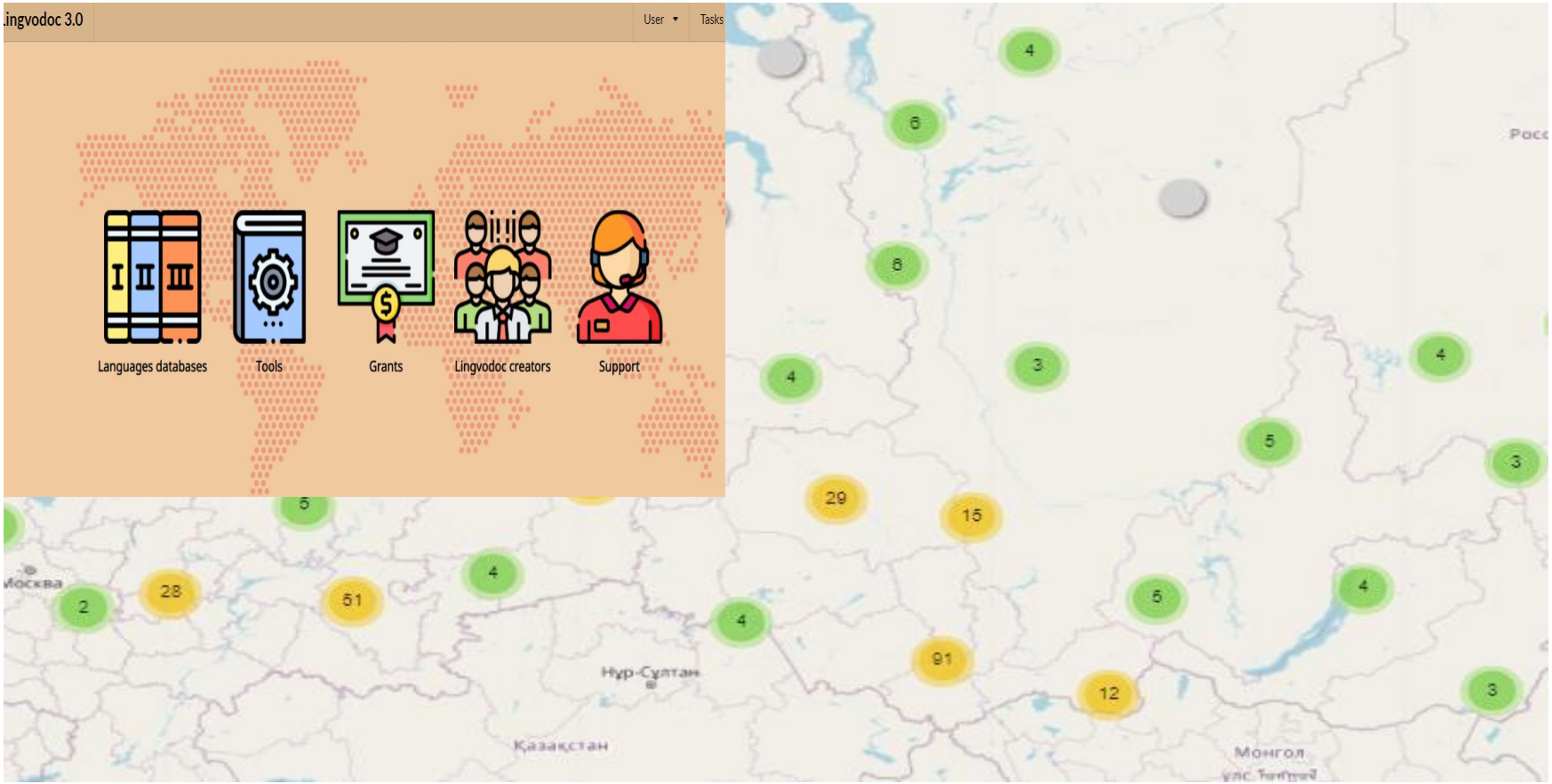


Яндекс Переводчик

8 languages

Bashkir
Mountmari
Mari
Russian
Tatar
Udmurt
Chuvash
Sakha

Linguistic Platform “Lingvodoc”



Turkic Morpheme Portal

Selected database language:

Tatar

Common part

Grammar

Thesaurus

Situations

Language part

Morphemes

Morphotactics

Situations in language

Multiword expressions

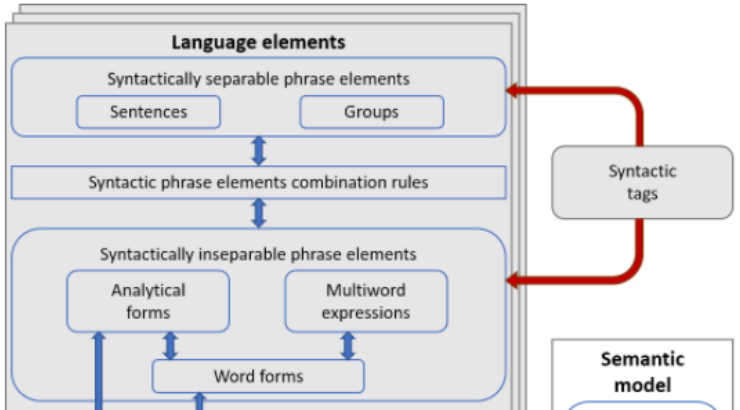
Tukic Morpheme Model

Development of scientific research in fields of turkology and agglutinative languages typology requires such software that take into account the structural and functional features of languages in question.

Turkic Morpheme web portal is a toolkit that takes into account core features of Turkic languages and meets the requirements for research activities in computational linguistics and typology.

This portal was created on the basis of the structural-parametric functional model of the Turkic morpheme and contains special linguistic databases that describe the categories of Turkic languages at different levels: morphological, syntactic, and semantic.

The portal can also be used in educational process as a reference system for Turkic languages.



Internet Portal “Turkic Morpheme”

http://modmorph.turklang.net

Functions of the linguistic portal

- Information and reference system
- Linguistic resource – linguistic database for solving applied tasks
- Collaboration platform
- A toolkit for creating terminology and designations unification
- Natural language processing code pipeline
- Research toolkit

Information and reference system: terminological base

Grammemes

Typological name (English)	Typological name (Russian)	National name (Tatar)
1-st person	1-е лицо	1-нче зат
2-st person	2-е лицо	2-нче зат
3-st person	3-е лицо	3-нче зат
Ablative	Исходный падеж	Чыгыш килеше
Accusative	Винительный падеж	Төшем килеше
Active	Основной залог	Төп юнәлеш
Adjective	Имя прилагательное	Сыйфат
Adverb	Наречие	Рәвеш

Information and reference system: terminological base

Typological name (English)	Interjection
Description and source of typological name (English)	Interjections are words that conventionally constitute utterances by themselves and express a speaker's current mental state or reaction toward an element in the linguistic or extralinguistic context (see Context, Communicative). Some English interjections are words such as yuk! 'I feel disgusted', ow! 'I feel sudden pain', wow! 'I feel surprised and I am impressed', aha! 'I now understand', hey! 'I want someone's attention', damn! 'I feel frustrated', and bother! 'I feel annoyed'. Such words are found in all languages of the world. Interjections may be defined using formal, semantic, or pragmatic criteria. From a formal point of view, an interjection is typically defined as a lexical form that (a) conventionally constitutes a nonelliptical utter- ance by itself, (b) does not enter into construction with other word classes, (c) does not take inflectional or derivational affixes, and (d) is monomorphemic. Ameka F. K. Interjections // Brown Keith (ed.) - Encyclopedia of Language and Linguistics. Elsevier, 2005. P. 743.
Grammatical category	Part of speech : Часть речи

Language part: Tatar

National name	Ымлык
Description and source of national name	Кешеләрнең хис-тойгыларын, эчке кичерешләрен һәм теләк-ихтыярын белдерә торган сүзләр ымлык дип атала. Ымлыклар лексик һәм грамматик яктан башка сүз төркемнәреннән нык аерылалар: 1) Ымлыклар эмоцияләрне һәм ихтыярны төшенчәләр аша чагылдырмыйлар, бәлки аларның турыдан-туры тәгъбирләре булып торалар. Әйтик, сокланү, гаҗәпләнү, курку, җирәнү сүзләре кешенең эмоцияләрен атап белдерсәләр, и! әй! ах! фу! ымлыклары саф эмоциональ сигналлар хезмәтен генә үтиләр. 2) Ымлыклар сирәк очракларда сүз ясагыч кушымчалар белән килсәләр дә, сүз төрләндергеч кушымчаларны кабул итмиләр; исемнәр кебек – сан, килеш, тартым белән һәм фигыльләр кебек зат, сан, заман, юнәлеш белән төрләнмиләр. 3) Ымлыклар, номинатив функция үтәмәгәнлектән, җөмлә кисәге була алмыйлар; алар я үзләре генә мөстәкыйль бер җөмлә төзиләр, я аның конструктив бер элементы сыйфатында кулланылалар.Мәсәлән: Т ф у ! 4) Ымлыклар экспрессив-эмоциональ төсмерләргә бай булалар, махсус интонация һәм кул-йөз хәрәкәтләре катнашлыгында әйтеләләр. // Татар грамматикасы: өч томда / проект җит. М.З. Зәкиев. – Тулыландырылган 2 нче басма. – Казан: ТӘһСИ, 2016. – Т. II. – 432 б. - С. 395.

Collaboration platform

Forum

General categories

News	1
Bug messages	0
General topics	0

Language categories

Language category: Altaic	0
Language category: Azerbaijani	0
Language category: Bashkir	0

Linguistic resource – linguistic database for solving applied tasks

Affixal morpheme

-ДАн : Ablative : Исходный падеж

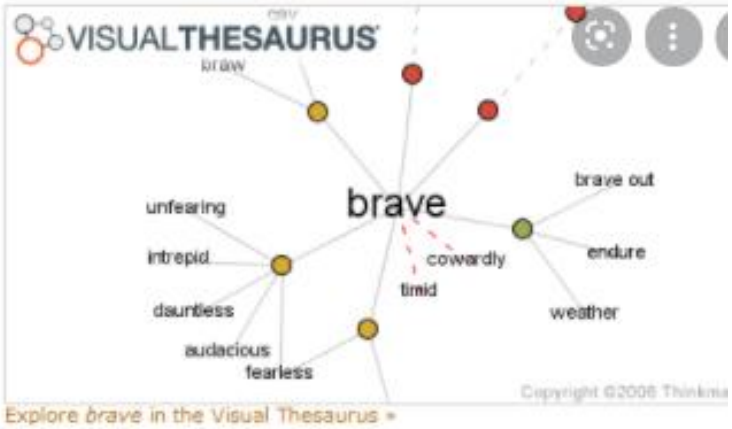
Full digital identifier	16.2.6.3
Name	-ДАн
Grammatical value	Ablative : Исходный падеж

Allomorphs

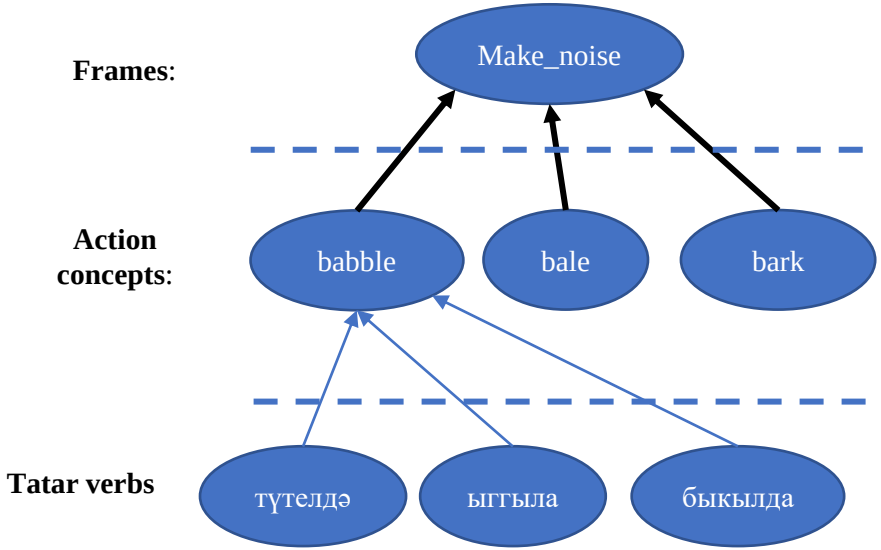
Name	Full digital identifier	Example	Translation
дан	16.2.6.3.1	абый-дан баерак	богаче брата
дән	16.2.6.3.2	тез-дән булды	стало по колено
нан	16.2.6.3.5	урман-нан кайтты	вернулся из леса
нән	16.2.6.3.6	идән-нән күтәрелде	поднялся с пола
тан	16.2.6.3.3	агач-тан ясалган	сделано из дерева
тән	16.2.6.3.4	биш-тән артык	больше пяти

Using a multilingual thesaurus in teaching:

Automatic creation of bilingual thematic dictionaries for any pair of languages



Knowledge Portal: Multilingual Turkic FrameNet



Make_noise : Шуметь

Name (Russian)	Шуметь
Name (English)	Make_noise

Root morphemes

Action concept	Root morpheme
babble : лопотать, лепетать	түтелдә
	ыггыла
	быкылда
	бытылда
	бытырда
	тәтелдә
	гыгылда
bale : тюкать	түкелдә
bark : лаять, гавкать, брехать, гавкнуть, залаять, пролаять, прогавкать	һаула
	һавылда

Roles in Frames: Damaging : повреждать

Roles in frames

Role	Description (Russian)	Description (English)	Is mandatory
Agent : Агент	Сознательная сущность, как правило, человек, который выполняет преднамеренное действие, которое приводит к повреждению Пациента.	The conscious entity, generally a person, that performs the intentional action that results in the damage to the Patien	✔
Cause : Причина	Событие, которое приводит к повреждению Пациента.	An event which leads to the damage of the Patient.	✔
Patient : Пациенс	Сущность, на которую воздействует Агент, так что она повреждается.	The entity which is affected by the Agent so that it is damaged.	✔
Character_of_end_state : Характер конечного состояния	Описание состояния Пациента после повреждения, включая описание тяжести и постоянства поврежденного состояния.	A description of the state of the Patient after the damaging has taken place, including descriptions of the severity and the permanence of the damaged condition.	
Degree : Мера	Степень повреждения Пациента.	The degree to which the Patient is damaged.	
Explanation : Объяснение	Состояние дел, на которое реагирует Агент при выполнении действия.	A state of affairs that the Agent is responding to in performing the action.	
Instrument : Инструмент	Субъект, управляемый Агентом, который взаимодействует с Пациентом для нанесения ущерба.	An entity directed by the Agent that interacts with a Patient to accomplish the damage.	

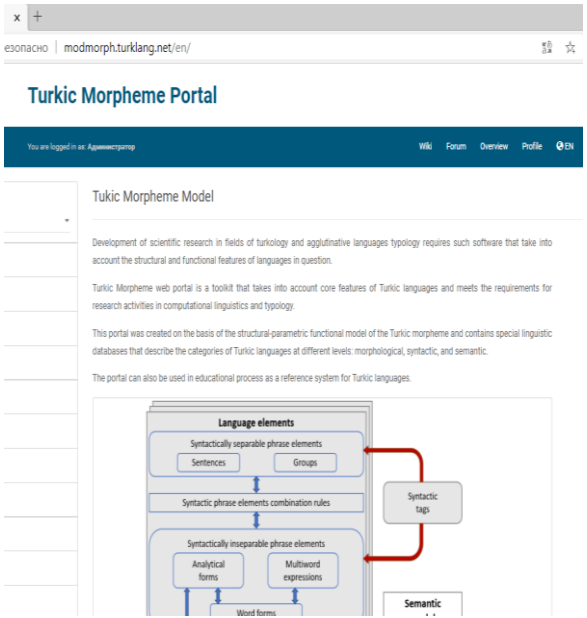
Status of the portal database

Current state of the database:

Number of languages	37
Number of grammar categories	19
Number of derivatives	15
Number of concept objects	18108
Number of action concepts	786
Number of object attribute concepts	2272
Number of action attribute Concepts	213
Number of affixal morphemes	1268
Number of affixal allomorphs	6462
Number of analytical morphemes	176
Number of root morphemes	104 256
Number of affixal rules for morphotactics	38401

Multilingual resource and interface

From Portal ⟶ to Computer Fund



Integration with GIS

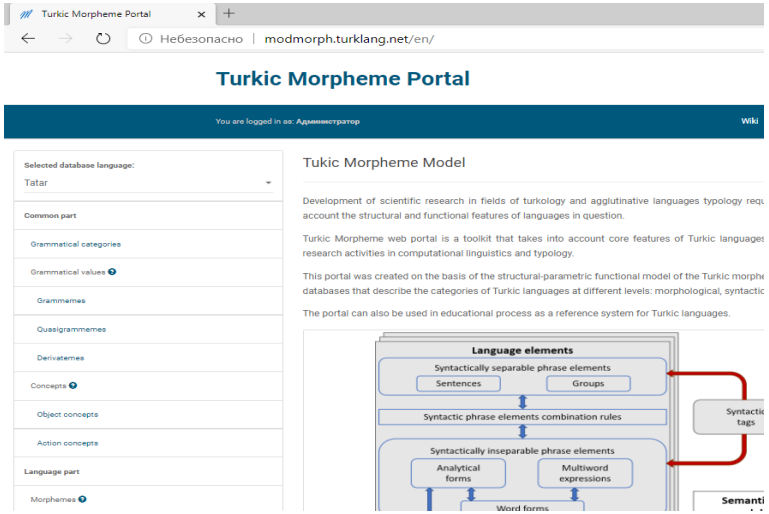
From Portal ⟶ to Computer Fund

Multilingual electronic corpus of Turkic languages

Corpus annotation: morphological, syntactic, semantic

Automatic database filling

Search for examples in the corpus



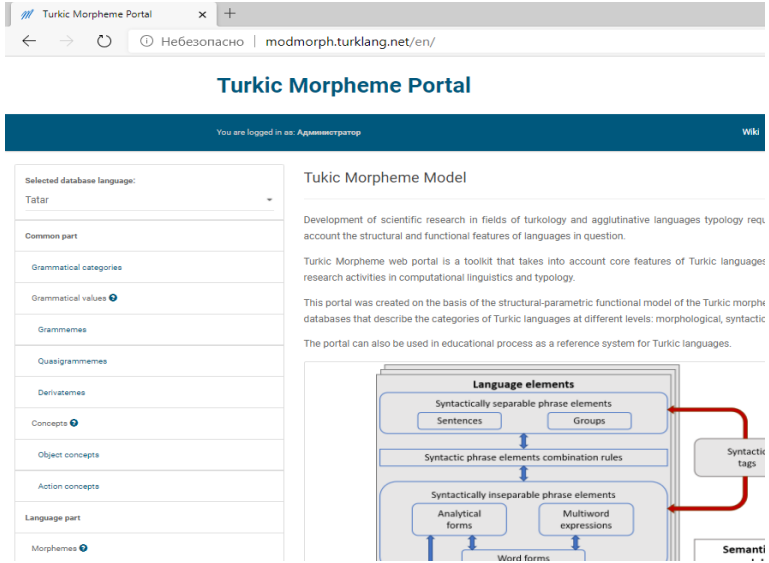
From Portal ⟶ to Computer Fund

Multilingual electronic corpus of Turkic languages
&
E-Learning system

Creation of domain ontology and database of question-answer models

Automatic generation of testing program on demand from a teacher

Creation of ICG (semantic grammars) for the knowledge control system



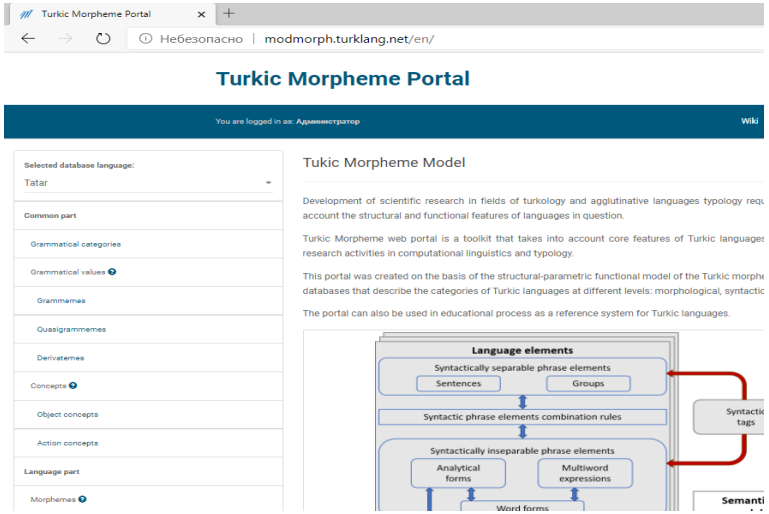
From Portal ⟶ to Computer Fund

Multilingual electronic corpus of Turkic languages
&
Subject ontologies

Generation of lexical and grammatical models (based on ontologies and annotated text databases)

Automatic generation of potentially viable grammatical models

Search for examples in databases of potential grammatical constructions



Recommendations

1. Developing language resources and tools from scratch is expensive. Nevertheless, such efforts can often be reduced by using existing resources and tools for structurally and lexically interconnected and more resource-intensive languages. In this regard, I would recommend that the conference draw the attention of the world scientific community and relevant motivated organizations to providing organizational and informational support to the creation and development of Open Multilingual Internet Platforms in order to adjust such resources and software for other languages. Multilingual Internet resources also contribute to the evolution of languages based on their comparative study and mutual lexical and grammatical enrichment, as well as bringing the specialists together and creating scientific teams in the process of designing and using the Internet Platform.

2. Research of the cognitive potential of Natural Languages and development of new technologies of Artificial Intelligence based on lexical and grammatical models of Natural Languages belong to the most demanded and breakthrough scientific and applied directions in the problems of creating explanatory artificial intelligence. In this regard, it is necessary to support and intensify the research and development in this area in order to create Artificial Intelligence systems with cognitive structures that ensure interaction with a person at the level of "understanding each other".

Thank you for attention!

Спасибо за внимание!

Игътибарыгыз өчен рәхмәт!

Назарларыңызға рахмет!

E'tiboringiz uchun rahmat!