

QARAQALPAQ TILI ELEKTRON KORPUSIN JARATIW TIYKARLARI

Sharbaev Jaras Bayrambay uli

Qaraqalpaq mámleketlik universiteti
Ámeliy filologiya kafedrası stajyor-oqıtıwshısı
Nókis, Qaraqalpaqstan, Ózbekstan.

Email: sharbaevj77@gmail.com

<https://doi.org/10.5281/zenodo.21843185>

Annotaciya. Bul maqalada qaraqalpaq tiliniń zamanagóy leksikasın óz ishine alǵan elektron korpustı jaratıw texnologiyaları hám onıń lingvistikalıq múmkinshilikleri talqılanadı.

Izertlewdiń baslı maqseti — qaraqalpaq tili resursların sanlastırw, tildiń statistikalıq hám morfologiyalıq parametrlerin analizlew ushın universal platforma jaratıw bolıp tabıladı.

Maqalada internet resursları, á sirese Wikipedia hám elektron GXQ tekstleri tiykarında jıynalǵan 2,5 millionnan artıq sóz qollanıwǵa iye bolǵan korpustıń arxitekturası kórsetilgen.

Shıǵıstaǵı tatar milliy korpustı (“Tugan tel”) sıyaqlı iri joybarlar tájiriybesine súylene otırıp, qaraqalpaq tili korpusında úsh basqıshlı izlew sisteması: morfologiyalıq, kontekstli (KWIC) hám statistikalıq (N-gramm) modulleri islep shıǵılmaqta.

Sonday-aq, maqalada korpus quramındaǵı tekstlerdi avtomatlastırılǵan morfologiyalıq teglew maseleleri hám tildiń agglyutinativlik qásiyetin esapqa alǵan halda izlew algoritmelerin optimallastırw usılları keltirilgen. Izertlew nátiyjeleri qaraqalpaq tilin oqıtıw, leksikografiya hám mashina awdarması sistemaların rawajlandırıwda zárúr derek bolıp xızmet etedi.

Gilt sózler: korpus, maǵlıwmatlar bazası, n-gramm, izlew sisteması, teglew, korpus quramı.

I. KIRISIW

Qaraqalpaq tiliniń elektron korpusın jaratıwda jáhán hám turkiy tiller korpus lingvistikası tarawındaǵı jetiskenlikler menen tanısıp ótiw zárúr. Atap aytqanda, usı waqıtqa shekem jaratılǵan hám dúnyaga belgili bolǵan turkiy jáne jáhán kólemindegi korpuslardan inglís¹, rus², qazaq³, ózbek⁴, tájik⁵ hám basqa da korpuslardı sanap ótsek boladı. Á sirese, Tatarstan Respublikası Ilimler akademiyası tárepinen jaratılǵan “Tugan tel” tatar tili milliy korpusı biz ushın úlgi retinde xızmet etiwı múmkin.

“Tugan tel” korpusı 200 millionǵa jaqın sózdi óz ishine alıp, tómendegi zárúr xızmetlerdi ámelge asırmaqta:

- **janrlar boyınsha quramı:** kórkem ádebiyat, GXQ, rásmiy hújjetler hám ilimiy baspalar teń salmaqlı jıynalǵan;

- **morfologiyalıq annotaciya:** PC-KIMMO texnologiyası tiykarında avtomatlastırılǵan morfologiyalıq teglew sisteması jolǵa qoyılǵan;

¹ <https://www.english-corpora.org/bnc/>

² <https://ruscorpora.ru/>

³ https://qazcorpus.kz/_tarihi-ishkorpus/index1.php

⁴ <https://uzcorpora.uz/>

⁵ <https://www.termcom.tj/>

• **metamağlıwmatlar:** hárbir teksttiń avtorı, janrı hám jazılğan waqtı haqqındaǵı tolıq baza jaratılǵan⁶.

Sonday-aq, dúnyadaǵı eń iri korpuslardan biri bolǵan Rus tili milliy korpusı⁷ (2 milliard sóz qollanıw) (RTMK) tájiriybesi de dıqqatqa ılayıq. RTMK diaxronik (XI ásirde XXI ásirge shekem) hám sociolingvistikalıq (ádebıy, awızeki, dialektlik) quramı boyınsha universal sistema esaplanadı.

RTMK platformasında tekstlerdi avtomatlastırılǵan morfologiyalıq hám semantikalıq teglew ushın **MyStem** hám **RuBik** sıyaqlı zamanagóy lingvistikalıq instrumentler qollanıladı⁸.

Bul tájiriybelerden kelip shıǵıp, biz jaratıp atırǵan Qaraqalpaq tili elektron korpusı da tómendegi tiykarǵı ilmiy-texnikalıq talaplarǵa juwap beriw maqset etildi:

1. Internet saytlarındaǵı tekstler: korpus tolıq qaraqalpaq tilindegi tekstlerden ibarat bolıp, ásirese internet resursları (Wikipedia, social tarmaqlar hám jańalıqlar saytları) tiykarında tildiń zamanagóy leksikasın sáwlelendirip beredi.

2. Kóp funkciyalı izlew sisteması: paydalanıwshılar maǵlıwmatlardı úsh túrli usılda izlew imkaniyatına iye boladı:

- **morfologiyalıq ózgesheligi boyınsha izlew:** Sózdiń grammatikalıq forması hám shaqapları boyınsha izlew;

- **kontekstli izlew (KWIC):** Sózdiń tekst ishindegi qollanıw ortalıǵın analiz qılıw ushın;

- **statistikalıq izlew (N-gramm):** 2-gramm, 3-gramm hám 4-gramm modelleri tiykarında tildiń n-gramm jiyiligin anıqlaw;

- **korpusaǵı barlıq jumsalǵan sózlerdiń dizimi:** korpusa jumsalǵan sózlerdiń barlıǵı sóz shaqabı sheńberinde korpusıń óz aldına bóliminde beriledi.

II. KORPUS USHÍN TEKSTLERDI TAYARLAW BASQÍSHÍ

Qaraqalpaq tiliniń elektron korpusın dúziw ushın Wikipediadaǵı 2025-jılǵa shekem bolǵan barlıq tekstler (10 000 nan aslam maqala yaki 2,5 mlnnan aslam token) bazası txt keńeytpeli formatta jıynaldı.

Bul baza boyınsha búgingi kúnde dáslepki tazalaw basqısı, tazalanǵan tekstlerdi bolsa morfologiyalıq teglew jumısları alıp barılmaqta.

Wikipediadaǵı bul materiallar mazmunı boyınsha tiykarınan hártúrli kategoriyalarǵa tiyisli bolıp, olardıń ishinde: *til, ádebiyat, geografiya, siyasat, huqıq, biologiya, anatomiya, geodeziya, sport, tariyx, zoologiya, IT, texnologiya, programmalastırıw h.t.b* janrlarǵa tiyisli tekstlerdiń bar ekenligi anıqlandı.

Tekstlerdi tazalaw basqışında keste, diagramma, SmartArt hám grafikalıq elementler, súwret, sızılmalar, HTML hám programmalastırıwǵa tiyisli basqa da tegler, kodlar alıp taslandı.

III. KORPUS MAǒLÍWMATLAR BAZASÍNÍN DÚZILISI

Elektron korpus maǒlıwmatlar bazasınıń dúzilisi tómendegi ierarxiyaǵa iye bolıp, onda eń úlken birlikten eń kishi birlikke qaray principi tiykarında is kóreledi.

⁶ <https://tugantel.tatar/>

⁷ <https://ruscorpora.ru/page/corpora-about>

⁸ N.Z.Abduraxmonova, Á.Z.Otemisov Kompyuter lingvistikasına kirisiw. Nókis – 2024

Bunda, eń úlken birlik tekst ishinde abzac, ekinshi dárejeli birlik gáp, úshinshi dárejeli birlik sóz, sóz bolsa óz ishinde tiykar hám affiks morfema dep belgilendi hám maǵlıwmatlar bazası qalıplestirildi:

```
<p>  
<phrase>  
<words>  
<iword> <!-- Tekst ishindegi sózdiń basqa kórinisi -->  
<item type="txt">sóz</item> <!-- Sózdiń grammatikalıq mánisi hám glossariysi-->  
<item type="gls">maǵlıwmat</item> <!-- Sózdiń sóz shaqabı (Part of Speech - POS) -->  
<item type="pos">n</item>  
</iword>  
</words>  
</phrase>  
</p>
```



1-súwret. Korpus maǵlıwmatlar bazası ierarxiyası

Tekstlerdi teglewde xalıqaralıq tájiriyyede keń qollanılauǵın, SIL International Group tárepinen islep shıǵılǵan, kem resurslı tillerdi dala ekspediciyaları arqalı úyreniwge, leksikalıq fondın sanlı islep shıǵıw hám korpusların jaratıwǵa arnalǵan FLEx⁹ (FieldWorks Language Explorer) programmasınan paydalanılmaqta.

Biraq, bunday keń kólemli jumislarda tek ǵana FLEx programmasınıń kúshi jetkiliksiz.

Házirgi waqıtta qaraqalpaq tilindegi avtomat annotator¹⁰lar ústinde jumislardıń alıp barılıp atırǵanı keleshekte korpus kólemin ele de jetilistiriwge úlken járdem beredi.

⁹ <https://software.sil.org/fieldworks/>

¹⁰ A. Otemisov and J. Sharbaev, "Development of a Karakalpak Morphological Analyzer in the Foma Environment," 2026 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBREIT), Yekaterinburg, Russian Federation, 2026, pp. 1-4, doi: 10.1109/USBREIT70063.2026.11580433. keywords: {Databases;Modeling;Tagging;Transducers;STEM;Printing;Foma;morphological analyzer;Karakalpak language;agglutination;FST;lexc;xfst;vowel harmony;morphology},

Korpustağı tekstlerdi morfologiyalıq teglew tártibi tómendegi kestedeki kórsetilgen terminlerge súyenedi:

Teg	Tolıq ataması	Qaraqalpaqsha ataması
n	noun	atlıq
adj	adjective	kelbetlik
num	numeral	sanlıq
v	verb	feyl
adv	adverb	ráwısh
pron	pronoun	almasıq
prep / post	preposition / postposition	kómekshi
conj	conjunction	dáneker
part	particle	tirkewish
interj	interjection	tańlaq

IV. JUWMAQ

Juwmaqlap aytqanda, qaraqalpaq tiliniń elektron korpusın jaratıw hám onı zamanagóy talaplar tiykarında rawajlandırıw — tilimizdiń sanlı dáwirdegi belsendiligin támiyinleytuǵın eń zárúrli ilmiy wazıypalardan biri bolıp tabıladı. Maqalada kórip shıǵılǵan arxitektura, atap aytqanda, 2,5 millionnan artıq sóz qollanılıwın óz ishine alǵan bazanı tayarlaw, ierarxiyalıq túrde dúziw hám úsh basqıshlı izlew sistemasın (morfologiyalıq, KWIC hám N-gramm) engiziw joybardıń ózek bólegin quraydı.

Alıp barılıp atırǵan bul izertlewler keleshegi keń bolıp, kelisiw basqıshlarında korpus quramın kórkem ádebiyat, publicistika hám awızeki sóylew tili materialları menen tolıqtırıw, sonday-aq avtomatlastırılǵan teglew modullerin qaraqalpaq tiliniń agglyutinativlik tabiyatına tolıq beyimlestiriw arqalı nátiyjelerdi bunnan da joqarı basqıshqa kóteriw múmkinshiligin beredi.

Usı korpusıń tolıq iske túsiwi qaraqalpaq til biliminde fundamental ilmiy izertlewlerdi jańa sapalıq dárejege kóterip qoymastan, al tildi oqıtıw metodikası, leksikografiya hám qosımsha programmalıq támiynatlar, atap aytqanda, mashina awdarması sistemaların jaratıwda sheshiwshi ilmiy-metodikalıq baza bolıp xızmet etedi.

Paydalanılǵan ádebiyatlar:

1. <https://tugantel.tatar/>
2. <https://ruscorpora.ru/page/corpora-about>
3. <https://software.sil.org/fieldworks/>
4. <https://qazcorpus.kz/tarihi-ishkorpus/index1.php>
5. <https://uzcorpora.uz/>
6. <https://www.termcom.tj/>
7. <https://tugantel.tatar/>
8. <https://ruscorpora.ru/page/corpora-about>

9. A. Otemisov and J. Sharbaev, "Development of a Karakalpak Morphological Analyzer in the Foma Environment," 2026 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBREIT), Yekaterinburg, Russian Federation, 2026, pp. 1-4, doi:10.1109/USBREIT70063.2026.11580433.
keywords:
{Databases;Modeling;Tagging;Transducers;STEM;Printing;Foma;morphological analyzer;Karakalpak language;agglutination;FST;lexc;xfst;vowel harmony;morphology},
<https://ieeexplore.ieee.org/document/11580433>.

10. Ótemisov A.Z., Sharbaev J.B Qaraqalpaq hám túrk tilleriniń parallel korpusın jaratıw máselesi, <https://ctcl.uz/index.php/uzmu/article/view/116>.

11. N.Z.Abduraxmonova, Á.Z.Otemisov Kompyuter lingvistikasına kirisiw. Nókis – 2024