ISNN 3030-3907

2026/3
VOLUME 8

DEVELOPMENT OF SCIENCE

ILMIY JURNAL



Development of science

Ilmiy jurnal

2026/3 VOLUME 8

ISSN 3030 -3907

Ilmiy jurnal OAK rayosatining 2024 yil 27 sentabrdagi 361-son qaroriga asosan 02.00.00 kimyo fanlari hamda 2025 yil 12 fevral 367- son qaroriga asosan 05.00.00 texnika fanlari bo‘yicha falsafa doktori (PhD) va fan doktori (DSc) ilmiy darajasiga talabgorlarning dissertatsiya ishlari yuzasidan dissertatsiyalari asosiy ilmiy natijalarini chop etish tavsiya etilgan milliy ilmiy nashrlar ro‘yxatiga kiritilgan.

BOSH MUHARRIR:
Temirov Alisher Hoshim o‘g‘li – texnika fanlari falsafa doktori (PhD),Dotsent.
TAHRIRIYAT KENGASH RAISI:
Fozilov Sadriddin Fayzullayevich – texnika fanlari doktori, professor
TAHRIRIYAT KENGASH A’ZOLARI

TEXNIKA FANLARI/ TECHNICAL SCIENCES/ ТЕХНИЧЕСКИЕ НАУКИ

Bozorov G‘ayrat Rashidovich
Buxoro Davlat Texnika Universiteti, professor
Abdurahmonov Olim Rustamovich
Buxoro Davlat Texnika Universiteti, professor
Adizov Bobur Zamirovich
O‘zbekiston Respublikasi Fanlar akademiyasi Umumiy va noorganik kimyo instituti, professor
Hayitov Ruslan Rustamjonovich
Buxoro Davlat Texnika Universiteti, professor
Jumayev Qayum Karimovich
Buxoro Davlat Texnika Universiteti, professor.
Berdiyev Dorob Murodovich
Islom Karimov nomidagi Toshkent Davlat Texnika Universiteti, professor
Maxmudov Rafiq Amonovich
Buxoro Davlat Texnika Universiteti, professor.
Tilloyev Lochin Ismatilloyevich
Buxoro Davlat Texnika Universiteti, dotsent.
Raxmonov Qahramon Sanakulovich
Buxoro Davlat Texnika Universiteti, dotsent.
Barakayev Nusratullo Rajabovich
Islom Karimov nomidagi Toshkent Davlat Texnika Universiteti, professor.
Jo‘rayev Xayrullo Fayziyevich
Buxoro Davlat Texnika Universiteti, professor.
Panoyev Erali Rajabboyevich
Buxoro Davlat Texnika Universiteti, dotsent.
Ataullayev Sherzod Nabiullyevich
Buxoro Davlat Texnika Universiteti, dotsent.
Murodov Malik Negmurodovich
Buxoro Davlat Texnika Universiteti, dotsent.
Axmedova Ozoda Bahronovna
Islom Karimov nomidagi Toshkent Davlat Texnika Universiteti.
Safarov Bahri Jumayevich
Buxoro Davlat Texnika Universiteti, dotsent.
Xaydarov Axtam Amonovich
Buxoro Davlat Texnika Universiteti, dotsent.
Safarov Jasur Alijon o‘g‘li
Buxoro Davlat Texnika Universiteti, dotsent.
Асадова Раъно Дилмуратовна
Тошкент Давлат техника университети доценти

KIMYO FANLARI/ CHEMICAL SCIENCES/ ХИМИЧЕСКИЕ НАУКИ

Axmedov Voxid Nizomovich
Buxoro Davlat Texnika Universiteti, professor
Umarov Baqo Bafoyevich
Buxoro Davlat Universiteti, professor
Kodirov Abduaxad Abduraximovich
Qarshi davlat Universiteti, professor
Qurbonov Mingniqul Jumagulovich
Qarshi davlat Universiteti, dotsent
Turemuratov Sharibay Naurizbaevich
O‘zbekiston Respublikasi Fanlar akademiyasi Qoraqalpog‘iston bo‘limi Bosh ilmiy kotibi, professor
Tursunov Kaxor Shonazarovich
Qarshi davlat Universiteti, professor
Do‘stov Hamro
Buxoro Davlat Texnika Universiteti, professor
Panoyev Nodir Shavkatovich
Buxoro Davlat Texnika Universiteti, dotsent.
Olimov Bobur Bahodir o‘g‘li
Buxoro Davlat Texnika Universiteti, dotsent.
Kasimov Sherzod Abduazimovich
Termiz Davlat Universiteti, professor.
Aliqulov Rustam Valiyevich
Termiz Davlat Universiteti, professor
Tropov Xamza Tursunovich
Samarqand Davlat Universiteti, professor

IQTISODIYOT FANLARI/ ECONOMICS/ ЭКОНОМИЧЕСКИЕ НАУКИ

Nizamov Aslitdin Badreddinovich
Buxoro Davlat Texnika Universiteti, professor.
Boboyev Akmal Choriyevich
Buxoro Davlat Texnika Universiteti, professor.
Ergashxojayeva Shaxnoza Djasurovna
Toshkent davlat iqtisodiyot universiteti, professor
Saidaxmedova Nodira Ilxomovna
Toshkent davlat iqtisodiyot universiteti, professor
Muxsinov Bekzod Toxirovich
Buxoro Davlat Texnika Universiteti, dotsent.
Shakarov Qulmat Ashirovich
Renessans ta'lim universiteti,dotsent.
Jumaeva Zulfiya Qayumovna
Buxoro Davlat Texnika Universiteti, dotsent.
Norova Salomat Yusupovna
Buxoro Davlat Texnika Universiteti, dotsent.
Avezova Shaxnoza Maxmudjonovna
Buxoro Davlat Texnika Universiteti, dotsent.
Jalilov Jamshid G‘anijonovich
Toshkent davlat iqtisodiyot universiteti, dotsent
Xasanova Gulruh Djumanazarovna
Buxoro Davlat Texnika Universiteti, dotsent.

PEDAGOGIKA FANLARI/ PEDAGOGICAL SCIENCES / ПЕДАГОГИЧЕСКИЕ НАУКИ

Siddiqova Sadoqat G‘afforovna
Buxoro Davlat Texnika Universiteti, dotsent.
Zoyirov Erkin Xalilovich
Buxoro Davlat Texnika Universiteti, DSc, professor
Gaffarov Feruz Hasanovich
Puchon Universiteti, *professor*
Ruziyev Davron Yuldashevich
Buxoro Davlat Universiteti, professor
Muradova Firuza Rashidovna
Buxoro Davlat Texnika Universiteti, professor.
Yunusova Gulandon Samiyevna
Buxoro Davlat Texnika Universiteti, professor.
To‘xtayeva Zebo Sharifovna
Buxoro Davlat Texnika Universiteti, professor.
Navro‘zova Gulchixra Nigmatovna
Buxoro Davlat Texnika Universiteti, dotsent.
Uzoqova Laylo Polvonovna
Buxoro muhandislik – texnologiya instituti, dotsent.
To‘raqulova Marjona Qiyom qizi
Buxoro Davlat Texnika Universiteti, dotsent.
Muradova Zarina Rashidovna
Buxoro Davlat Texnika Universiteti, dotsent.
Murodov Sanjar Aslonovich
Buxoro Davlat Texnika Universiteti, dotsent.
Shodiyev Jahongir Jo‘raqulovich
Buxoro Davlat Texnika Universiteti, dotsent
Gaffarov Laziz Hasanovich
Buxoro muhandislik – texnologiya instituti, dotsent.

Tahririyat a’zolarning to‘liq ro‘yxati jurnalning www.devos.uz saytida keltirilgan.

Tahririyat hay’ati a’zolarining tarkibini o‘zgartirish hamda qo‘shimcha kiritish yoki tahririyat a’zoligidan chiqarish bosh muharrir tomonidan amalga oshiriladi.

ISSN 3030-3907

Development of science ilmiy jurnali 2024 yil 15 aprel kuni №256445 sonli guvohnoma bilan Davlat ro‘yxatidan o‘tgan.

To‘plam elektron ko‘rinishda (pdf formatida) mualliflarga taqdim etiladi. To‘plamga kiritiladigan maqolalarning mazmuni, ulardagi statistik ma’lumotlar va me’yoriy hujjatlarning to‘g‘riligiga hamda har qanday fikr - mulohazalarga mualliflarning o‘zlari mas’ul hisoblanadilar. Talablarga javob bermaydigan maqolalar to‘plamga kiritilmaydi. Tashkiliy qo‘mita maqola matnini qisqartirish, qisman tuzatish kiritish, seksiyalarga joylashtirish huquqiga ega.

ISSN 3030-3907

QORAQALPOQ TILI MATNLARINI AVTOMATIK QAYTA ISHLASHNING HISOBLASH USULLARI VA MODELLARI

Abdullayev Sherzod Shavkatjonovich
Oriental universiteti dotsenti
E-mail: abdulla.sherzod87@gmail.com
Allaberdiyeva Durdana Guranmuratovna
Raqamli texnologiyalar va sun'iy intellektni rivojlantirish ilmiy-tadqiqot instituti doktoranti
E-mail: durdonallaberdiyeva39@gmail.com

Annotatsiya. Mazkur maqolada qoraqalpoq tili uchun tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasining hozirgi holati tizimli ravishda tahlil qilinadi. Qoraqalpoq tili turkiy tillar oilasining qipchoq-no‘g‘ay guruhiga mansub, qariyb 800 ming–2 million kishi so‘zlashadigan, YuNESKO tasnifi bo‘yicha “zaif” (vulnerable) maqomidagi til bo‘lib, hisoblash lingvistikasi nuqtai nazaridan “kam resursli” (low-resource) tillar qatoriga kiradi. Tadqiqotda tilning agglutinativ morfologiyasi, unli singarmonizmi va ko‘p yozuvli (lotin, kirill, arab) tabiati keltirib chiqaradigan asosiy hisoblash muammolari ko‘rib chiqiladi. So‘nggi yillarda yaratilgan korpuslar (dilmash, KAASC), qoidaga asoslangan morfologik tahlilgichlar (Apertium/FST), neyron mashina tarjimasi modellari (NLLB) hamda yaqin qardosh tillardan (o‘zbek, qozoq) bilim ko‘chirish (cross-lingual transfer) yondashuvlari qiyosiy baholanadi. Tahlil natijalari kam resursli sharoitda eng samarali yo‘nalish sifatida ko‘p tilli oldindan o‘rgatilgan modellarni qardosh tillar yordamida moslashtirish ekanligini ko‘rsatadi. Maqola so‘ngida soha rivoji uchun amaliy tavsiyalar va kelajakdagi tadqiqot yo‘nalishlari taklif etiladi.

Kalit so‘zlar: qoraqalpoq tili, tabiiy tilni qayta ishlash, kam resursli tillar, turkiy tillar, morfologik tahlil, neyron mashina tarjimasi, bilim ko‘chirish, til korpusi, agglutinativ morfologiya, NLLB, LLM, sub-so‘z tokenizatsiyasi.

Abstract. This paper presents a systematic analysis of the current state of Natural Language Processing (NLP) for the Karakalpak language. Karakalpak belongs to the Kipchak–Nogai group of the Turkic language family and is spoken by roughly 800,000 to two million people; classified as "vulnerable" by UNESCO, it falls among the "low-resource" languages from a computational-linguistics standpoint. The study examines the main computational challenges that arise from the language's agglutinative morphology, vowel harmony, and multi-script nature (Latin, Cyrillic, and Arabic). Recently developed corpora (dilmash, KAASC), rule-based morphological analyzers (Apertium/FST), neural machine translation models (NLLB), and cross-lingual transfer approaches from closely related languages (Uzbek, Kazakh) are evaluated comparatively. The findings indicate

ISSN 3030-3907

that, under low-resource conditions, the most effective direction is adapting multilingual pretrained models with the help of related languages. The paper concludes with practical recommendations for advancing the field and directions for future research.

Keywords: Karakalpak language, natural language processing, low-resource languages, Turkic languages, morphological analysis, neural machine translation, transfer learning, language corpus, agglutinative morphology, NLLB, LLM, subword tokenization.

Аннотация. В настоящей статье проводится систематический анализ современного состояния обработки естественного языка (Natural Language Processing, NLP) для каракалпакского языка. Каракалпакский язык относится к кипчакско-ногайской группе тюркской языковой семьи, на нём говорят приблизительно от 800 тысяч до двух миллионов человек; согласно классификации ЮНЕСКО он имеет статус «уязвимого» (vulnerable) и с точки зрения компьютерной лингвистики входит в число «малоресурсных» (low-resource) языков. В исследовании рассматриваются основные вычислительные проблемы, обусловленные агглютинативной морфологией языка, гармонией гласных (сингармонизмом) и его многописьменным характером (латиница, кириллица, арабское письмо). Сравнительной оценке подвергаются созданные в последние годы корпуса (dilmash, KAASC), морфологические анализаторы на основе правил (Apertium/FST), модели нейронного машинного перевода (NLLB), а также подходы переноса знаний (cross-lingual transfer) из близкородственных языков (узбекского, казахского). Результаты анализа показывают, что в условиях ограниченных ресурсов наиболее эффективным направлением является адаптация многоязычных предобученных моделей с помощью родственных языков. В завершение статьи предлагаются практические рекомендации для развития данной области и направления дальнейших исследований.

Ключевые слова: каракалпакский язык, обработка естественного языка, малоресурсные языки, тюркские языки, морфологический анализ, нейронный машинный перевод, перенос знаний, языковой корпус, агглютинативная морфология, NLLB, LLM, субсловная токенизация.

1. KIRISH.

So‘nggi o‘n yillikda chuqur o‘rganish (deep learning) va transformer arxitekturasiga asoslangan til modellari tabiiy tilni qayta ishlash sohasini tubdan o‘zgartirdi. Biroq bu yutuqlar dunyodagi tillar o‘rtasida juda notekis taqsimlangan: zamonaviy NLP tizimlarining sezilarli qismi ingliz, xitoy, ispan kabi katta hajmdagi raqamli matn resursiga ega bo‘lgan tillarga yo‘naltirilgan. Dunyodagi

ISSN 3030-3907

yetti mingdan ortiq tilning aksariyati esa “kam resursli” toifaga kiradi — ya‘ni ular uchun yetarli hajmdagi annotatsiyalangan korpuslar, leksik bazalar va tayyor lingvistik vositalar mavjud emas. Bunday tillar uchun zamonaviy til texnologiyalarini yaratish nafaqat texnik, balki ijtimoiy-madaniy ahamiyatga ega masala hisoblanadi, chunki raqamli muhitda til mavjudligi uning kelajakdagi hayotiyligi bilan bevosita bog‘liq.

Qoraqalpoq tili ana shunday kam resursli tillarning yorqin namunasidir. U turkiy tillar oilasining qipchoq tapмog‘iga, aniqrog‘i qipchoq-no‘g‘ay kichik guruhiga mansub bo‘lib, qozoq va no‘g‘ay tillariga eng yaqin turadi [1]. Til asosan O‘zbekiston tarkibidagi Qoraqalpog‘iston Respublikasida tarqalgan. Ona tili sifatida unda so‘zlashuvchilar soni turli manbalarda 872 mingdan (2023) qariyb ikki milliongacha baholanadi [2]. Shunga qaramay, hozirgi kunda yetakchi tijoriy tarjima platformalari, jumladan keng tarqalgan avtomatik tarjima xizmatlari ushbu tilni qo‘llab-quvvatlamaydi, bu esa tilning raqamli infratuzilmasini mustaqil ravishda shakllantirish zaruriyatini yanada kuchaytiradi [3].

Mazkur maqolaning maqsadi qoraqalpoq tili uchun NLP sohasidagi mavjud ilmiy-amaliy ishlanmalarni tizimli tarzda umumlashtirish, ulardan foydalanilgan asosiy uslublarni qiyosiy baholash va istiqbolli yo‘nalishlarni belgilashdan iborat. Tadqiqotda quyidagi vazifalar hal etiladi: birinchidan, qoraqalpoq tilining NLP nuqtai nazaridan muhim bo‘lgan lingvistik xususiyatlari tavsiflanadi; ikkinchidan, mavjud til resurslari va korpuslari sanab o‘tiladi; uchinchidan, morfologik tahlil, tokenizatsiya, mashina tarjimasi va bilim ko‘chirish kabi asosiy uslubiy yondashuvlar tahlil qilinadi; to‘rtinchidan, ushbu yondashuvlarning samaradorligi qiyosiy baholanib, kelajakdagi ishlar uchun tavsiyalar ishlab chiqiladi.

Maqola sharh-tahliliy (review) xarakteriga ega bo‘lib, unda asosan ochiq ilmiy manbalar, retsenziyalangan maqolalar hamda ochiq kodli loyihalar natijalariga tayanilgan. Bu yondashuv kam o‘rganilgan sohada mavjud bilimlarni bir tizimga keltirish va kelgusi tadqiqotchilar uchun metodologik asos yaratish imkonini beradi.

2. Qoraqalpoq tilining lingvistik tavsifi

Qoraqalpoq tilining NLP uchun muhim bo‘lgan tipologik xususiyatlarini tushunish ushbu til uchun hisoblash modellarini to‘g‘ri loyihalashda hal qiluvchi ahamiyatga ega. Quyida tilning asosiy lingvistik parametrlari umumlashtirilgan (1-jadval).

ISSN 3030-3907

1-jadval.

Qoraqalpoq tilining asosiy lingvistik va sotsiolingvistik parametrlari

Parametr	Qiymati
Til oilasi	Turkiy > Umumturkiy > Qipchoq > Qipchoq–no‘g‘ay
Eng yaqin tillar	Qozoq, no‘g‘ay
ISO 639-3 kodi	kaa
So‘zlashuvchilar	~872 000 (2023) – ~2 mln (turli manbalarda)
Hudud	Qoraqalpog‘iston Respublikasi (O‘zbekiston)
Yozuv tizimlari	Lotin (joriy rasmiy), kirill, arab (tarixiy)
Morfologik tip	Agglutinativ, unli singarmonizmiga ega
So‘z tartibi	SOV (ega–to‘ldiruvchi–kesim)
Dialektlar	Shimoli-sharqiy va janubi-g‘arbiy
YuNESKO maqomi	Zaif (vulnerable)

Manba: [1], [2] asosida muallif tomonidan umumlashtirildi.

2.1. Agglutinativ morfologiya

Boshqa turkiy tillar singari qoraqalpoq tili ham agglutinativ tuzilishga ega: so‘z negiziga ketma-ket qo‘shiluvchi affikslar yordamida grammatik ma‘nolar ifodalanadi. Bitta o‘zakdan ko‘plab so‘zshakllar hosil bo‘ladi; masalan, otlarga egalik, kelishik va son qo‘shimchalari, fe‘lga esa nisbat, mayl, zamon, shaxs-son ko‘rsatkichlari birin-ketin qo‘shiladi. Natijada bir o‘zak yuzlab, ba‘zan minglab turli so‘zshaklga ega bo‘lishi mumkin. Bu hodisa NLP uchun jiddiy oqibatlarga olib keladi: lug‘at hajmining keskin kengayishi (vocabulary explosion), so‘zshakllarning siyrak taqsimlanishi (data sparsity) va so‘zni so‘ddiy bo‘laklarga ajratish (tokenizatsiya) muammosining murakkablashuvi kuzatiladi. Shu sababli morfologik segmentatsiya va sub-so‘z darajasidagi modellashtirish bunday tillar uchun markaziy ahamiyat kasb etadi.

2.2. Unli singarmonizmi va fonologiya

Qoraqalpoq tilida unli singarmonizmi amal qiladi: so‘z tarkibidagi unlilar til oldi/orqasi va lablanish bo‘yicha o‘zaro moslashadi, shu sababli ko‘pchilik affikslar bir necha fonetik variantga ega bo‘ladi. Bu hodisa morfologik tahlilgichlar va matnni sintezlash tizimlarini loyihalashda hisobga olinishi shart bo‘lgan qoidaviy murakkablikni keltirib chiqaradi, biroq ayni paytda affikslarning variantlari oldindan aytib bo‘luvchi bo‘lganligi sababli qoidaga asoslangan modellashtirishni qulaylashtiradi.

ISSN 3030-3907

2.3. Yozuv tizimlarining ko‘pligi

Qoraqalpoq tili XX–XXI asrlar davomida arab, lotin va kirill yozuvlaridan foydalangan; hozirgi kunda rasmiy ravishda lotin grafikasiga asoslangan alifboga o‘tilmoqda. Mavjud raqamli matnlarning sezilarli qismi kirill yozuvida saqlanib qolgani sababli, korpuslarni yagona standartga keltirishda ishonchli transliteratsiya (yozuvlararo o‘girish) muhim bosqichga aylanadi. Bundan tashqari, qoraqalpoq matnlarida o‘zbek va rus tillari bilan kod almashinuvi (code-switching) hamda imloviy nomuvofiqliklar tez-tez uchraydi, bu esa matnni normallashtirish (normalization) vazifasini murakkablashtiradi.

3. Asosiy hisoblash muammolari va chaqiriqlar

Yuqorida keltirilgan lingvistik xususiyatlar va resurs tanqisligi qoraqalpoq tili uchun NLP tizimlarini yaratishda bir qator o‘ziga xos chaqiriqlarni keltirib chiqaradi. Ularni quyidagi guruhlarga ajratish mumkin:

- **Ma‘lumotlar tanqisligi.** Yetarli hajmdagi yorliqlangan (annotated) korpuslar, parallel matnlar va leksik bazalarning yo‘qligi nazoratli o‘rgatish (supervised learning) uslublarini bevosita qo‘llashni cheklaydi.
- **Morfologik murakkablik.** So‘zshakllarning ko‘pligi tufayli yuzaga keladigan lug‘at portlashi va siyraklik standart so‘z darajasidagi modellarning samaradorligini pasaytiradi.
- **Standartlashuv muammolari.** Yozuv tizimlarining ko‘pligi va imloviy variativlik tufayli matnlarni yagona ko‘rinishga keltirish zarurati tug‘iladi.
- **Baholash vositalarining yetishmasligi.** Tilga xos standartlashtirilgan sinov to‘plamlari (benchmark) va etalon ma‘lumotlar (gold-standard) yaqin-yaqingacha deyarli mavjud emas edi, bu modellarni xolisona qiyoslashni qiyinlashtirardi.
- **Mutaxassis va infratuzilma tanqisligi.** Til bo‘yicha hisoblash lingvistikasi sohasidagi tadqiqotchilar soni cheklangan, shuningdek tayyor ochiq kodli vositalar yaqin vaqtgacha kam edi.

Ushbu chaqiriqlar bir-biri bilan o‘zaro bog‘liq: masalan, ma‘lumotlar tanqisligi baholash vositalarining yetishmasligini kuchaytiradi, morfologik murakkablik esa mavjud kichik korpuslardan unumli foydalanishni yanada qiyinlashtiradi. Shu bois zamonaviy yondashuvlar ko‘pincha bir necha muammoni bir vaqtda hal etishga qaratilgan kompleks yechimlar shaklida taklif etiladi.

4. Til resurslari va korpuslar

So‘nggi yillarda qoraqalpoq tili uchun bir qator muhim raqamli resurslar yaratildi. Garchi ularning hajmi katta resursli tillar bilan solishtirganda hamon cheklangan bo‘lsa-da, bu to‘plamlar tilning NLP infratuzilmasi uchun poydevor vazifasini o‘taydi. Quyidagi 2-jadvalda asosiy resurslar umumlashtirilgan.

ISSN 3030-3907

2-jadval.

Qoraqalpoq tili uchun mavjud asosiy NLP resurslari

Resurs	Turi	Tavsifi
dilmash (Tahrirchi)	Parallel korpus	~300 000 jumla jufti: o‘zbek–qoraqalpoq, rus–qoraqalpoq, ingliz–qoraqalpoq yo‘nalishlari uchun [3]
FLORES+ devtest (kaa)	Baholash to‘plami	1 012 jumladan iborat qoraqalpoqchaga tarjima qilingan etalon sinov to‘plami [3]
KAASC	Monolingual korpus	23 ta maktab darsligidan tuzilgan korpus; TF-IDF asosida to‘xtatuvchi so‘zlar ro‘yxati yaratilgan [4]
Yirik monolingual korpuslar	Monolingual korpus	Lotin yozuvidagi millionlab jumlalardan iborat, til modellarini o‘rgatishga mo‘ljallangan ochiq to‘plamlar
Gatitos leksikoni	Leksik baza	Kam resursli tillar uchun ko‘ptilli lug‘at; ~4 000 ingliz–qoraqalpoq jufti [5]

Manba: [3], [4], [5] asosida tuzildi.

Resurslarni yaratishda qo‘llanilgan uslublar ham diqqatga sazovor. Masalan, parallel jumlalarni saralashda til-agnostik jumla vektorlari (LaBSE — Language-agnostic BERT Sentence Embedding) yordamida moslik darajasini hisoblash usuli qo‘llanilgan, bu esa bir kitob yoki hujjatning turli tilli tarjimalaridan sifatli jumla juftliklarini avtomatik ajratib olish imkonini bergan [3], [6]. Bunday yondashuv kam resursli tillar uchun parallel korpus yig‘ishning amaliy va kengaytiriluvchan usuli hisoblanadi.

5. Metodologik yondashuvlar

Qoraqalpoq tili uchun NLP masalalarini hal etishda qo‘llaniladigan yondashuvlarni shartli ravishda uch katta yo‘nalishga ajratish mumkin: (1) qoidaga asoslangan lingvistik usullar; (2) statistik va neyron uslublar; (3) qardosh tillardan bilim ko‘chirishga tayanuvchi ko‘ptilli yondashuvlar. Quyida har biri batafsil ko‘rib chiqiladi.

5.1. Qoidaga asoslangan morfologik tahlil

Agglutinativ tillar uchun an‘anaviy va hamon dolzarb bo‘lgan yondashuv — chekli holatli o‘tkazgichlar (finite-state transducers, FST) yordamida morfologik tahlilgich qurishdir. Bu modellar tilning morfotaktikasini (affikslarning birikish qoidalarini) va leksik bazani (o‘zaklar lug‘atini) yagona avtomatga kompilatsiya qiladi hamda so‘zni o‘zak va affikslarga ajratish (analiz) hamda aksincha so‘zshakl

ISSN 3030-3907

hosil qilish (generatsiya) vazifalarini ikki yo‘nalishda bajaradi. Ochiq kodli Apertium platformasi va lttoolbox kabi vositalar aynan shunday FST-larni yaratishga mo‘ljallangan bo‘lib, ko‘plab turkiy tillar uchun morfologik tavsiflar shu asosda ishlab chiqilgan [7].

FST yondashuvining afzalliklari — katta hajmdagi o‘rgatuvchi ma‘lumotni talab qilmasligi, yuqori tezligi, ixchamligi va natijalarning lingvistik jihatdan izohlanuvchanligidir. Kamchiligi esa malakali lingvist mehnatini va batafsil qoidalar to‘plamini talab qilishi, shuningdek til normasidan chetga chiqishlar (xato imlo, jonli nutq) bilan yomon kurashishidir. Shu sababli zamonaviy tadqiqotlarda FST natijalarini neyron modellar bilan birlashtiruvchi gibrid arxitekturalar istiqbolli deb topilmoqda: bunday tizimda FST nomzod tahlillarni keltirib chiqaradi, neyron model esa kontekst asosida eng ehtimoliy variantni tanlaydi yoki FST tanimagan so‘zni mustaqil tahlil qiladi. O‘xshash gibrid yechim yaqin qardosh qozoq tili uchun KazMorphCorpus-2025 doirasida muvaffaqiyatli sinovdan o‘tkazilgan va u qoraqalpoq tiliga moslashtirishga nomzod metodologiya sifatida xizmat qilishi mumkin [8].

5.2. Tokenizatsiya va sub-so‘z darajasidagi modellashtirish

Agglutinativ morfologiya keltirib chiqaradigan lug‘at portlashi muammosini hal etishning eng samarali zamonaviy usuli — sub-so‘z (subword) tokenizatsiyasidir. Bunda matn butun so‘zlar emas, balki tez-tez uchraydigan belgilar ketma-ketliklariga ajratiladi, natijada noyob so‘zshakllar ham ma‘no tashuvchi kichik bo‘laklar orqali ifodalanadi. Bayt-juftlikni kodlash (Byte-Pair Encoding, BPE) va SentencePiece kabi til-agnostik algoritmlar aynan shu maqsadda keng qo‘llaniladi [9], [10]. Bu usullar turkiy tillar uchun ayniqsa qo‘l keladi, chunki ular morfologik chegaralarni statistik tarzda taxminan qamrab oladi va siyraklik muammosini sezilarli darajada yumshatadi. Aynan SentencePiece tokenizatori qoraqalpoq mashina tarjimasi modellarida ham qo‘llanilgan [3].

5.3. Neyron mashina tarjimasi

Qoraqalpoq tili uchun NLP sohasidagi eng katta amaliy yutuq mashina tarjimasi yo‘nalishida qo‘lga kiritilgan. Transformer arxitekturasi [11] zamonaviy neyron mashina tarjimasining (Neural Machine Translation, NMT) asosini tashkil etadi. Kam resursli sharoitda noldan model o‘rgatish samarasiz bo‘lgani sababli, ustun yondashuv — oldindan ko‘p tilda o‘rgatilgan yirik modelni maqsadli til ma‘lumotlari bilan moslashtirish (fine-tuning) hisoblanadi. “Birorta ham til ortda qolmasin” (No Language Left Behind, NLLB) loyihasining ko‘ptilli modeli aynan shunday tayanch sifatida xizmat qiladi [12].

Mamasaidov va Shopulatov (2024) qoraqalpoq tili uchun NLLB modelini yuqorida tilga olingan dilmash parallel korpusi yordamida moslashtirib, o‘zbek, rus

ISSN 3030-3907

va ingliz tillari bilan ikki yo‘nalishli tarjima uchun ochiq neyron modellarni yaratdi. Tadqiqotchilar turli model variantlari va o‘rgatish strategiyalarini qiyoslab, mavjud bazaviy ko‘rsatkichlardan ustun natijalarga erishdilar; modellar va ma‘lumotlar to‘plamlari ommaviy foydalanish uchun ochiq e‘lon qilindi [3]. Bu ish Open Language Data Initiative (OLDI) doirasidagi umumiy vazifa sifatida amalga oshirilib, kam resursli tillar uchun “korpus + moslashtirilgan model + baholash to‘plami” ko‘rinishidagi yaxlit metodologiyaning namunaviy ko‘rsatkichi bo‘ldi.

5.4. Qardosh tillardan bilim ko‘chirish (cross-lingual transfer)

Kam resursli tillar uchun eng istiqbolli strategiyalardan biri — morfologik va leksik jihatdan yaqin bo‘lgan katta resursli qardosh tillardan bilim ko‘chirishdir. Qoraqalpoq tili qozoq va no‘g‘ay tillariga, shuningdek geografik va aloqaviy jihatdan o‘zbek tiliga yaqin bo‘lgani uchun bu yondashuv tabiiy ravishda qo‘l keladi. XLM-R kabi ko‘ptilli oldindan o‘rgatilgan kodlovchilar [13] va turkiy tillar uchun maxsus interlingua yondashuvlari [14] bilim ko‘chirish uchun asos yaratadi.

2026-yilda e‘lon qilingan TurkicNLP ochiq kodli vositalar to‘plami bu yo‘nalishdagi yaqqol misoldir. U turkiy tillar uchun yagona, izchil NLP konveyerini (tokenizatsiya, morfologik tahlil, so‘z turkumlarini belgilash, sintaktik tahlil, nomli obyektlarni aniqlash, yozuvlararo transliteratsiya va boshqalar) taqdim etadi hamda qoidaga asoslangan FST va neyron modellarni shaffof tarzda birlashtiradi [15]. Diqqatga sazovor jihati shundaki, qoraqalpoq tili uchun maxsus o‘rgatuvchi ma‘lumot bo‘lmasa-da, o‘zbek tilini “proksi” (vositachi) sifatida ishlatuvchi vektorlar orqali noldan (zero-shot) sintaktik tahlil amalga oshirilgan: hisobotga ko‘ra, bu yondashuv qoraqalpoq tili uchun so‘z turkumlarini belgilashda 82,1% UPOS aniqligiga va 51,0 LAS ko‘rsatkichiga erishgan [15]. Bu natija hech qanday maqsadli til ma‘lumotisiz ham qardosh tillar yordamida amaliy darajadagi tahlil mumkinligini ko‘rsatadi.

5.5. Katta til modellari va baholash to‘plamlari

Katta til modellari (Large Language Models, LLM) davrida kam resursli tillar uchun yangi imkoniyatlar va shu bilan birga yangi muammolar paydo bo‘ldi. Bir tomondan, ko‘ptilli LLM-lar qoraqalpoqcha matnlarni ma‘lum darajada qayta ishlay oladi; boshqa tomondan, ularning bu tildagi ishonchliligi va aniqligi yetarlicha o‘rganilmagan. Shu munosabat bilan turkiy tillar uchun mahalliy (native) tilni tushunish bo‘yicha baholash to‘plamlari, jumladan TUMLU benchmarki ishlab chiqilmoqda; bunday vositalar ko‘ptilli modellarni adolatli va madaniy kontekstni hisobga olgan holda baholash imkonini beradi [16]. Qoraqalpoq tilini bunday benchmarklarga to‘liq kiritish kelajakdagi muhim vazifalardan biridir.

ISSN 3030-3907

Shuningdek, MADLAD-400 kabi yirik, ko‘ptilli va hujjat darajasidagi auditdan o‘tgan korpuslar [17] kam resursli tillar uchun monolingual ma‘lumot manbai sifatida xizmat qiladi va ulardan til modellarini oldindan o‘rgatishda foydalanish mumkin. Standartlashtirilgan baholashda esa BLEU kabi an‘anaviy o‘lchovlar bilan birga, ularning cheklovlarini hisobga olgan holda chrF va boshqa belgi darajasidagi metrikalar tavsiya etiladi [18].

6. Qiyosiy tahlil va muhokama

Yuqorida ko‘rib chiqilgan yondashuvlarni samaradorlik, ma‘lumotga bo‘lgan talab va lingvistik izohlanuvchanlik mezonlari bo‘yicha qiyoslash mumkin (3-jadval). Har bir yondashuvning o‘z kuchli va zaif tomonlari mavjud bo‘lib, amalda ularning kombinatsiyasi eng yaxshi natija beradi.

3-jadval.

NLP yondashuvlarining qiyosiy tahlili

Yondashuv	Ma‘lumotga talab	Izohlanuvchanlik	Kam resursga moslik
Qoidaga asoslangan FST	Past (lingvist mehnati)	Yuqori	Yuqori
Noldan neyron model	Juda yuqori	Past	Past
Ko‘ptilli modelni moslashtirish (fine-tuning)	O‘rtacha	Past–o‘rtacha	Yuqori
Qardosh tildan bilim ko‘chirish (zero-shot)	Juda past	O‘rtacha	Juda yuqori
Gibrid (FST + neyron)	O‘rtacha	Yuqori	Yuqori

Manba: muallif tahlili.

Qiyosiy tahlil shuni ko‘rsatadiki, kam resursli sharoitda noldan neyron model qurish eng kam samarali strategiya hisoblanadi, chunki u juda katta hajmdagi ma‘lumotni talab qiladi. Aksincha, ko‘ptilli oldindan o‘rgatilgan modellarni moslashtirish va qardosh tillardan bilim ko‘chirish kam ma‘lumot bilan ham amaliy natija beradi. Qoidaga asoslangan FST esa morfologik tahlil kabi izohlanuvchanlik muhim bo‘lgan vazifalarda ustun bo‘lib qoladi. Shuning uchun amaliyotda eng maqbul yo‘l — ushbu yondashuvlarni birlashtiruvchi gibrid

ISSN 3030-3907

arxitekturalardir: morfologik daraja uchun FST, semantik va sintaktik darajalar uchun esa ko‘ptilli neyron modellar.

Shuningdek, qoraqalpoq tilining o‘zbek va qozoq tillariga yaqinligi alohida strategik ustunlik beradi. O‘zbek tili uchun ishlab chiqilgan resurslar va vositalar — jumladan Universal Dependencies formatidagi treebanklar va morfologik tahlilgichlar [19] — qoraqalpoq tiliga moslashtirish uchun qimmatli boshlang‘ich nuqta bo‘lib xizmat qilishi mumkin. Bu o‘zaro yaqinlikdan tizimli foydalanish kelgusi tadqiqotlarning markaziy g‘oyasi bo‘lishi lozim.

7. Kelajak istiqbollari va tavsiyalar

O‘tkazilgan tahlil asosida qoraqalpoq tili uchun NLP sohasini rivojlantirishda quyidagi ustuvor yo‘nalishlar va amaliy tavsiyalarni ajratib ko‘rsatish mumkin:

- **Yorliqlangan korpuslarni kengaytirish.** So‘z turkumlari, morfologik belgilar va sintaktik bog‘lanishlar bilan annotatsiyalangan Universal Dependencies formatidagi treebankni yaratish ustuvor vazifa bo‘lishi kerak.
- **Gibrid morfologik tahlilgich.** Apertium asosidagi FST tahlilgichini neyron disambiguatsiya moduli bilan birlashtirgan ochiq vositani ishlab chiqish tavsiya etiladi.
- **Qardosh tildan transferdan tizimli foydalanish.** O‘zbek va qozoq tillari resurslari va modellarini vositachi sifatida ishlatish hamda zero-shot/few-shot yondashuvlarni kengaytirish maqsadga muvofiq.
- **Baholash infratuzilmasi.** Tilga xos standartlashtirilgan benchmark va etalon to‘plamlarni (TUMLU kabi tashabbuslarga qoraqalpoq tilini kiritish orqali) yaratish modellarni xolisona qiyoslash uchun zarur.
- **Nutq texnologiyalari.** Avtomatik nutqni aniqlash (ASR) va matndan nutq sintezi (TTS) yo‘nalishlari hozircha kam o‘rganilgan bo‘lib, kelgusida alohida e‘tibor talab qiladi.
- **Ochiq fan tamoyillari.** Resurslar va modellarni ochiq litsenziya ostida e‘lon qilish (dilmash misolidagidek) jamoaviy rivojlanishni tezlashtiradi.

Bu yo‘nalishlar bir-birini to‘ldiradi: korpuslarning kengayishi gibrid modellar sifatini oshiradi, baholash infratuzilmasi esa erishilgan natijalarni o‘lchash va taqqoslash imkonini beradi. Eng muhimi, qoraqalpoq tili NLP-si yakka tartibdagi loyihalar emas, balki turkiy tillar uchun umumiy ekotizimning bir qismi sifatida rivojlantirilishi maqsadga muvofiqdir.

8. XULOSA.

Mazkur maqolada qoraqalpoq tili uchun tabiiy tilni qayta ishlash sohasining hozirgi holati tizimli tahlil qilindi. Tadqiqot shuni ko‘rsatdiki, qoraqalpoq tili kam resursli, agglutinativ va ko‘p yozuvli til sifatida NLP nuqtai nazaridan o‘ziga xos

ISSN 3030-3907

murakkabliklarni keltirib chiqaradi: ma‘lumotlar tanqisligi, morfologik portlash, standartlashuv muammolari va baholash vositalarining yetishmasligi asosiy chaqiriqlardir.

Shu bilan birga, so‘nggi yillarda erishilgan natijalar — dilmash parallel korpusi va FLORES+ baholash to‘plamining yaratilishi, NLLB asosidagi neyron tarjima modellarining moslashtirilishi, TurkicNLP doirasida o‘zbek tili orqali zero-shot sintaktik tahlilning namoyish etilishi — tilning raqamli infratuzilmasi jadal shakllanayotganidan dalolat beradi. Qiyosiy tahlil eng samarali strategiya sifatida ko‘ptilli oldindan o‘rgatilgan modellarni qardosh tillar yordamida moslashtirish hamda qoidaga asoslangan va neyron usullarni birlashtiruvchi gibrid arxitekturalarni ko‘rsatdi.

Kelgusi tadqiqotlar yorliqlangan korpuslarni kengaytirish, gibrid morfologik tahlilgichlar yaratish, qardosh tillardan tizimli transfer va tilga xos baholash infratuzilmasini shakllantirishga qaratilishi lozim. Qoraqalpoq tili NLP-sini turkiy tillar uchun umumiy ochiq ekotizimning bir qismi sifatida rivojlantirish nafaqat ilmiy, balki tilning raqamli davrda hayotiyligini ta‘minlash uchun ham strategik ahamiyatga ega.

FOYDALANILGAN ADABIYOTLAR

1. Berdimuratov E., Dáwletov A. Házirgi qaraqalpaq tili. — Qaraqalpaqstan, 1979. (oliy o‘quv yurtlari filologiya talabalari uchun darslik).
2. Karakalpak language. // Ethnologue (2024); Wikipedia: Karakalpak language. — URL: https://en.wikipedia.org/wiki/Karakalpak_language
3. Mamasaidov M., Shopulatov A. Open Language Data Initiative: Advancing Low-Resource Machine Translation for Karakalpak. — arXiv:2409.04269, 2024. — URL: https://arxiv.org/abs/2409.04269
4. Dataset of Karakalpak language stop words (KAASC). // Data in Brief, 2023. — URL: https://www.sciencedirect.com/science/article/pii/S2352340923002305
5. Jones A., Caswell I., Firat O., Saxena I. GATITOS: Using a New Multilingual Lexicon for Low-Resource Machine Translation. // Proceedings of EMNLP 2023. — P. 371–405.
6. Feng F., Yang Y., Cer D., Arivazhagan N., Wang W. Language-agnostic BERT Sentence Embedding (LaBSE). — arXiv:2007.01852, 2020.
7. Forcada M.L., Ginestí-Rosell M., Nordfalk J. et al. Apertium: a free/open-source platform for rule-based machine translation. // Machine Translation, 2011. — Vol. 25. — P. 127–144.
8. Hybrid AI architectures for automatic text correction in the Kazakh language (KazMorphCorpus-2025). // PMC, 2025. — URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12741073/

ISSN 3030-3907

9. Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units (BPE). // Proceedings of ACL 2016. — P. 1715–1725.
10. Kudo T., Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. — arXiv:1808.06226, 2018.
11. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need. // Advances in Neural Information Processing Systems (NeurIPS), 2017.
12. NLLB Team, Costa-jussà M.R., Cross J. et al. No Language Left Behind: Scaling Human-Centered Machine Translation. — arXiv:2207.04672, 2022.
13. Conneau A., Khandelwal K., Goyal N. et al. Unsupervised Cross-lingual Representation Learning at Scale (XLM-R). // Proceedings of ACL 2020. — P. 8440–8451.
14. Mirzakhalov J. Turkic Interlingua: A Case Study of Machine Translation in Low-Resource Languages. — Master's thesis, University of South Florida, 2021.
15. TurkicNLP: An NLP Toolkit for Turkic Languages. — arXiv:2602.19174, 2026. — URL: https://github.com/turkic-nlp/turkicnlp
16. TUMLU: A Unified and Native Language Understanding Benchmark for Turkic Languages. — 2025.
17. Kudugunta S., Caswell I., Zhang B. et al. MADLAD-400: A Multilingual and Document-Level Large Audited Dataset. // Advances in Neural Information Processing Systems (NeurIPS), 2024.
18. Post M. A Call for Clarity in Reporting BLEU Scores. // Proceedings of WMT 2018. — P. 186–191.
19. Universal Dependencies Treebank for Uzbek; de Marneffe M.-C., Manning C., Nivre J., Zeman D. Universal Dependencies. // Computational Linguistics, 2021. — Vol. 47(2). — P. 255–308.

ISSN 3030-3907