

TURKIY TILLAR TAJRIBASI ASOSIDA QORAQALPOQ TILI UCHUN MORFOLOGIK KORPUS YARATISH

Esemuratov Abdimurat Esemuratovich

Filologiya fanlari bo'yicha falsafa doktori (PhD), O'zbekiston san'at va madaniyat
instituti Nukus filiali, mustaqil tadqiqotchi

Raushan Aymbetova

Qoraqalpoq tili va adabiyoti o'qituvchisi, Qoraqalpoq akademik litseyi, IIV,
O'zbekiston Respublikasi

<https://doi.org/10.5281/zenodo.18951664>

Annotatsiya: Ushbu maqola qoraqalpoq tili uchun morfologik korpus va teglangan modellarni yaratish muammosiga bag'ishlangan. Tadqiqotda turkiy tillar (o'zbek, qozoq, turk) uchun mavjud morfologik resurslar tahlil qilinib, ularni qoraqalpoq tiliga moslashtirish imkoniyatlari ko'rib chiqiladi. Transfer o'rganish usuli orqali mavjud turkiy tillar modellarini moslashtirish konsepsiyasi taklif etiladi.

Kalit so'zlar: qoraqalpoq tili, morfologik tahlil, korpus lingvistikasi, turkiy tillar, agglyutinatativ tillar, NLP, transfer o'rganish.

KIRISH

So'nggi o'n yilliklarda NLP texnologiyalarining jadal rivojlanishi lingvistik landshaftni tubdan o'zgartirdi. Biroq, qoraqalpoq tili kam resursli tillar qatorida ushbu yutuqlardan chetda qolmoqda. Bu til turkiy tillarning qipchoq guruhiga mansub bo'lib, agglyutinatativ tuzilishga ega. So'z yasovchi qo'shimchalarning boy tizimi maxsus resurslar – morfologik korpuslar va avtomatik analizatorlarni yaratishni talab qiladi. Hozirda qoraqalpoq tili bo'yicha to'liq raqamli morfologik ma'lumotlar bazasi mavjud emas, bu esa lingvistik tadqiqotlar va amaliy dasturlarni ishlab chiqishni sezilarli susaytirmoqda.

O'zbek, qozoq va turk tillari kabi qardosh tillar rivojlangan morfologik modellar va korpuslarga ega bo'lib, bu texnologiyalar transferi uchun yo'l ochadi.

ADABIYOTLAR TAHLILI

Veitsman va Hartmann (2025) Markaziy Osiyo turkiy tillari uchun NLP rivojlanish darajasini tahlil qilib, kuchli nomutanosiblikni aniqlagan: qozoq va o'zbek tillari resurslar ekotizimiga ega bo'lsa, qoraqalpoq tili jiddiy orqada qolmoqda. Asosiy to'siqlar – rus tilining ustunligi, zaif internet infratuzilmasi va keng ko'lamli tashabbuslarning yo'qligi [1].

Tukeyev U. va boshq. (2020) turkiy tillar uchun FST asosidagi morfologik segmentatsiya metodikasini taklif qilib, o'zbek tilida 94-95% aniqlikka erishgan. Bu yondashuv cheklangan ma'lumotlar sharoitida neyron modellarni samarali o'qitish imkonini beradi [2].

Çöltekin (2022) turk tili uchun resurslarni – muvozanatli korpuslar, treebank-resurslar, NER to'plamlari, parallel korpuslarni tavsiflagan. Muallif resurslarning litsenziyalar bilan cheklanganligi va og'zaki nutq uchun sifatli to'plamlar yetishmasligini qayd etgan [3].

Isbarov va boshq. (2025) sakkiz turkiy tilda (jumladan qoraqalpoq) TUMLU benchmarkini yaratib, LLM-modellarning tushunish sifatini taqqoslagan. Natijalar kam resursli tillar va milliy alifbolar o'rtasidagi qiyinchiliklarni ko'rsatgan [4].

Turganbaeva (2022) qoraqalpoq tilida beshta asosiy so'z yasash usulini ajratgan: affiksatsiya, kompozitsiya, leksik-semantik, leksik-sintaktik va fonetik usullar [5].

Yazar va boshq. (2025) turk tilining diaxronik o'zgarishlarini tahlil qiluvchi «Turkronicles» korpusini (45,375 hujjat, 842 mln so'z) yaratgan [6].

Surrey Morphology Group (2019–2025) qozoq, qoraqalpoq va o'zbek tillarining qiyosiy-morfosintaktik tadqiqotini olib bormoqda [7].

Otemisov va Esemuratov (2024) qoraqalpoq tilini raqamlashtirish muammolarini tahlil qilib, kompleks yondashuv zarurligini ko'rsatgan [8].

TRANSFER O'RGANISH KONSEPSIYASI

Turkiy tillarning umumiy lingvistik asosga ega ekanligi fonetik, morfologik, leksik va sintaktik darajalarda transfer imkoniyatlarini yaratadi.

Tukeyev (2020) tomonidan o'zbek tili uchun ishlab chiqilgan FST analizator qoraqalpoq tili uchun eng yaqin moslama bo'lishi mumkin. Moslashtirish uchun qo'shimchalar ro'yxatini qayta ko'rib chiqish, qoraqalpoq tiliga xos elementlarni qo'shish va fonetik o'zgarishlar qoidalarini ishlab chiqish talab etiladi.

Çöltekin (2022) tavsiflagan turk tili resurslari teglar tizimini ishlab chiqishda namuna bo'lishi, TUMLU benchmark esa modellarni sinovdan o'tkazishda muhim ahamiyatga ega.

MORFOLOGIK KORPUS YARATISH METODOLOGIYASI

Qoraqalpoq tili uchun morfologik korpus yaratish quyidagi bosqichlarni o'z ichiga oladi:

1. Matn materiallarini to'plash: badiiy adabiyot, rasmiy hujjatlar, internet resurslari va og'zaki nutq transkripsiyalari.
2. Teglash tizimini ishlab chiqish: so'z turkumi, grammatik kategoriyalar (son, kelishik, egalik, zamon, shaxs, nisbat), so'z yasaliş qo'shimchalari.
3. FST analizatorni moslashtirish: qo'shimchalar bazasini qayta ko'rib chiqish, fonetik o'zgarishlarni hisobga olish, morfonologik qoidalarini qo'shish.
4. Ekspert tekshiruvi: avtomatik teglangan matnlarni lingvist-ekspertlar tomonidan tekshirish.
5. Sinovdan o'tkazish: TUMLU benchmark yordamida korpus va modellar sifatini baholash.

XULOSA

Tadqiqot qoraqalpoq tilining agglyutinatib tabiati avtomatik morfologik tahlilga alohida talablar qo'yishini ko'rsatadi. FST va neyron tarmoq yondashuvlari cheklangan ma'lumotlar sharoitida ham – transfer o'rganishni to'g'ri qo'llash orqali – samarali bo'lishi mumkin. Keyingi qadamlar teg sxemalarini rasmiylashtirish, ma'lumotlar to'plamlarini kengaytirish va lingvistlar, dasturchilar hamda pedagoglar o'rtasida fanlararo hamkorlikni rivojlantirishni o'z ichiga oladi.

Adabiyotlar, References, Литературы:

1. Veitsman, Y., & Hartmann, M. (2025). Recent advancements and challenges of Turkic Central Asian language processing. In Proceedings of the Workshop on NLP for Low-resource Languages (LoResLM 2025). Association for Computational Linguistics. – P. 309–324.
2. Tukeyev U., Karibayeva A., Zhumanov Zh. Morphological segmentation method for Turkic language neural machine translation // Cogent Engineering. – 2020. – Vol. 7, No. 1. – Article 1856500.
3. Çöltekin, Ç. (2022). Resources for Turkish natural language processing. Natural Language Engineering, 28(4), 543–566.
4. Isbarov, J., Akhundjanova, A., Hajili, M., Huseynova, K., Gaynullin, D., Rzayev, A., ... & Ataman, D. (2025). TUMLU: A Unified and Native Language Understanding Benchmark for

Turkic Languages. Proceedings of ACL 2025, 22816–22838.

5. Turganbaeva, P. N. (2022). Ways of Word Formation in the Karakalpak Language. Indiana Journal of Multidisciplinary Research, 2(1), 11–13.
6. Yazar T., Kutlu M., Bayırlı İ. K. Turkronicles: Diachronic resources for the fast evolving Turkish language // Language Resources and Evaluation. – 2025. – Vol. 59. – P. 1–33. – DOI: 10.1007/s10579-025-09857-w.
7. Surrey Morphology Group. (2019–2025). Morpho-syntax of mutual intelligibility in the Turkic languages of Central Asia [Research project]. University of Surrey. <https://www.smg.surrey.ac.uk/projects/mutual-intelligibility/>
8. Otemisov, A. Z., & Esemuratov, A. E. (2024). The need to digitize Karakalpak language: problems and solutions. Models and Methods in Modern Science, 3(11), 15–18. DOI: 10.5281/zenodo.12670228.