QORAQALPOQ TILI MORFOANNOTATORINI YARATISH ASOSLARI

O‘temisov Aziz Zarliqbaevich,
PhD, dotsent
utemisov.aziz@mail.ru
Qoraqalpoq davlat universiteti

Annotatsiya. Maqolada qoraqalpoq korpus lingvistikasi sohasida muhim masala bo‘lgan avtomatik morfologik annotatsiyalovchi dastur yaratish, bunda, FST, Foma (chekli avtomatik transdyuserlar)lardan unumli foydalanish haqida so‘z boradi. Maqolada qoraqalpoq tili morfologiyasini avtomatlashtirishning metodologik asosi va dastlabki prototipi berib o‘tilgan.

Kalit so‘zlar: *morfologiya, FST, Foma, xfst, lexc, chekli avtomat transdyuserlari, NLP, POS, Teg, Python, interfeys.*

Abstract. The article discusses the development of an automatic morphological annotation program, which is a crucial issue in the field of Karakalpak corpus linguistics. It explores the efficient use of Finite-State Transducers (FST) and the Foma compiler for this purpose. The methodological foundations and an initial prototype for automating Karakalpak language morphology are presented.

Keywords: *morphology, FST, Foma, xfst, lexc, finite-state transducers, NLP, POS, Tag, Python, interface.*

Morfologik tahlilda FST va Foma texnologiyalari

Tabiiy tilni qayta ishlash (NLP) va korpus lingvistikasida agglyutinativ tillar morfologiyasini modellashtirishning eng samarali usuli chekli holat o‘zgartirgichlari (Finite-State Transducers - FST) hisoblanadi[1].

FST – bu ikki darajali (lexical and surface) ko‘rinishlarni o‘zaro bog‘lovchi matematik model bo‘lib, u so‘z shakllarini tahlil va sintez qilish imkonini beradi. FST texnologiyasi so‘z asosi va unga qo‘shiladigan affikslar zanjirini grafik

ko‘rinishda tasvirlash orqali tilning morfologik va fonologik qoidalarini yuqori aniqlikda raqamlashtiradi[2].

Foma – bu FST tamoyillariga asoslangan, ochiq kodli va yuqori samaradorlikka ega dasturiy vosita bo‘lib, Mans Hulden tomonidan ishlab chiqilgan[3:29-32]. U lexc (leksikonlarni aniqlash) va xfst (fonologik qoidalarni yozish va h.k.) tillarini qo‘llab-quvvatlaydi. Fomaning asosiy afzalligi shundaki, u chegaraviy holat avtomatlarni minimallashtirish va kompyuter xotirasida juda kichik o‘rin egallagan holda millionlab so‘z shakllarini bir necha millisekundlarda tahlil qilish qobiliyati hisoblanadi[4:820-827]. Bugungi kunda jahon tajribasida kam resursli va murakkab morfologiyaga ega tillar uchun avtomatik annotatorlar yaratishda Foma eng ishonchli vosita sifatida qo‘llanilmoqda.

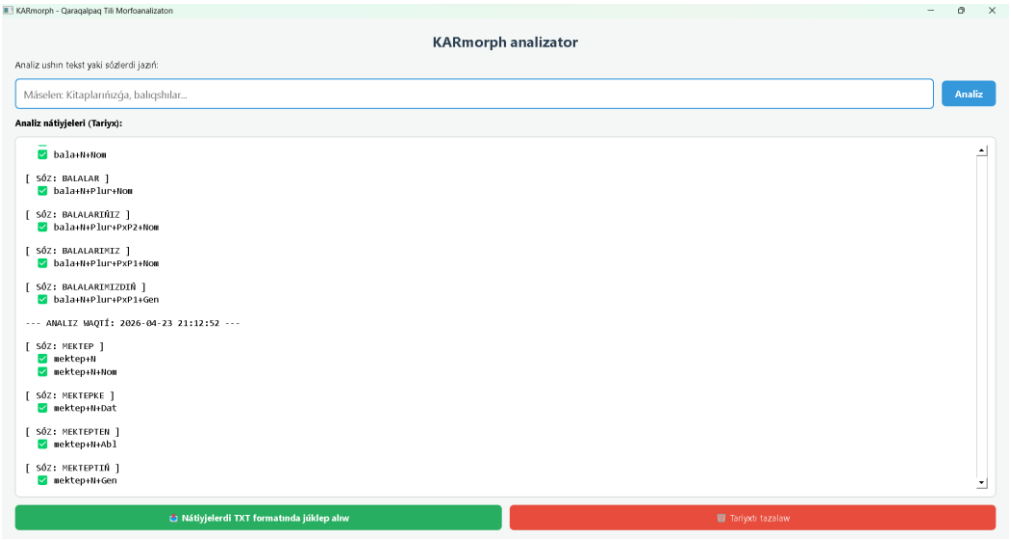
Morfologik annotatsiya tizimini standartlashtirish masalasida Britto va boshqalar ta’kidlaganidek, tizim nafaqat zamonaviy til qatlamlarini, balki tarixiy matnlarni ham qamrab olishi uchun bir vaqtning o‘zida ham "keng (broad), ham "aniq" (strict) bo‘lishi kerak. Mualliflarning fikricha, bunday muvozanatga erishish uchun teglarni iyerarxik darajalarga (Parts of Speech, Inflectional tags) ajratish va ularni tanlashda sintaktik distributiv mezonlarga tayanish zarur. Bizning qoraqalpoq tili uchun ishlab chiqilgan modelimiz ham xuddi shu prinsipga asoslanadi: agglyutinativlik til tabiatidan kelib chiqib, birinchi bosqichda so‘z turkumi (POS), keyingi bosqichlarda esa unga birikkan morfologik ko‘rsatkichlar (kelishik, shaxs-son va h.k.) ketma-ketlikda annotatsiyalanadi. Bu yondashuv korpus tarkibidagi matnlarning xronologik davriga qaramasdan, izchil va bir xil tahlil qilinishini kafolatlaydi[5].

Morfologik jihatdan boy tillar tizimini avtomatik tahlil qilishda o‘ziga xos qiyinchiliklar mavjud bo‘lib, bu jarayon tilga xos bo‘lgan morfemalar va so‘z yasash qoidalarini (o‘zakka qo‘shimcha qo‘shilishi, fonetik o‘zgarishlar: assimilyatsiya, unlilar uyg‘unligi va h.k.) chuqur o‘rganishni talab etadi. Durst va boshqalarning

aytishicha, o‘rganuvchilar korpusini tahlil qilishda xatolarni teglash sxemasi aynan o‘sha tilning strukturaviy xususiyatlaridan kelib chiqib ishlab chiqilishi zarur[6:39-54]. Venger tili misolida olib borilgan tadqiqotlar shuni ko‘rsatadiki, ona tili sohiblari uchun yaratilgan avtomatik morfologik vositalarni (masalan, magyarlanc) o‘rganuvchilar matnlariga qo‘llash orqali tildagi o‘zgarishlarni statistik tahlil qilish mumkin. Bu jarayonda to‘rtta asosiy omil: so‘z o‘zagi allomorflari, affiksatsiyadagi fonologik o‘zgarishlar, unlilar uyg‘unligi va affikslarning qat’iy tartibi markaziy o‘rinni egallaydi.

Mualliflarning fikricha, bunday avtomatlashtirilgan tizimlar nafaqat lingvistik tadqiqotlar, balki tillarni chet tili sifatida o‘qitish metodikasini takomillashtirish, xususan, o‘quv lug‘atlari va darsliklarni yaratish uchun ham juda muhim manba hisoblanadi. Shunday qilib, morfologik analizatorlar yordamida olingan ma’lumotlar til o‘rganish jarayonidagi tizimli xatolarni aniqlash va ularni bartaraf etish strategiyalarini ishlab chiqish imkonini beradi.

Qoraqalpoq tili uchun yaratilayotgan KARmorph morfoanalizatori bugungi kunda dastlabki test sinovlarida sinab ko‘rilmoqda. Hozirgi vaqtda uning leksikon bazasini hozirgi holatidan ham to‘ldirish, tilning har bir qatlamiga tegishli so‘zlar guruhini e’tibordan chetda qoldirmaslik, morfotaktik qoidalarni yanada to‘ldirish va eng muhimi omonimiya hodisasining yechimini topish ustida ishlar bajarilmoqda. Analizatorning natijalari bilan quyidagi rasmda tanishishingiz mumkin:



1-rasm. Qoraqalpoq tili KARmorph analizatori interfeysi va analiz natijasi

Qoraqalpoq tili uchun morfoanalizator yaratishda eng avvalo turkiy tillar bo‘yicha shu yo‘nalishda o‘rganilgan ishlar bilan tanishdi. Xususan Chag‘ri Choltekinning turk[9], N. A. Israilova va P. S. Bakasovaning qirg‘iz[10:100-116] va R.A.Gilmullin, B.T.Ganeev va M.Z.Suleymanovlarning tatar[11:114-122] tillari uchun FST modellarida yaratilgan keng imkoniyatlarni o‘zida mujassam etgan analizatorlarni ham ta’kidlash joiz.

Biz maqolamizda so‘z yuritayotgan analizatorning Foma muhiti bilan integratsiyalashgan python dasturlash tilida yozilgan interfeysi mavjud bo‘lib, u matn kiritish oynasi, tahlil tugmasi, natija oynasi, natijalarni txt kengaytmada yuklash imkoniyati va tarixni tozalash kabi bir qancha qismlarni o‘z ichiga oladi. Matn kiritish oynasiga berilgan so‘rovning natijasi agar u to‘g‘ri bo‘lsa yashil stikerlarda, noma’lum bo‘lsa qizil stikerlarda va unknown ma’lumoti bilan qaytariladi. Hozircha analizatorning desktop versiyasi takomillashtirilmoqda, buning asosiy sababi test natijalarining qay darajada sifatli ekanini kuzatish imkoniyati hisoblanadi. Test sinovlari to‘liq yakunlanib, analizator belgilangan standartlarga javob beradigan darajaga olib chiqish rejalashtirilgan.

1-jadval. Foma muhitida leksikon bazasida foydalanilgan teglar royxati

So‘z turkumlari / Kategoriya	Teg (Tag)	Izoh	Misollar
Ot (Noun)	+N	Sodda otlar (tub so‘zlar)	*bala, kitap, ilim*
	+N_COMP	Murakkab, qo‘shma va juft otlar	*aq quw, awılkeńes, es-aqıl*
Atoqli otlar	+NP_loc	Toponimlar (jer-suw atamaları)	*Angliya, Moskva, Nókis*
	+NPperson	Odam nomlari	*Abat, Aziz, Gúlnaz*
Fel (Verb)	+V	Tub fe’llar	*bar, kel, jaz, oqı*
	+V_COMP	Qo‘shma fe’llar	*baqsıshılıq et, qarawıllıq qıl*
Ravish (Adverb)	+Adv_Cont	Ravishlar	*tez, áste, dárhal*
Sifat (Adjective)	+Adj_INFL	Sodda sifatlar	*aq, qara, biyik, uzın*
	+Adj_COMP	Qo‘shma sifatlar	*sarı ala, ashıq jasıl*
Olmosh (Pronoun)	+Pron_Pers	Betlik olmoshlari	*men, sen, ol*
	+Pron_Refl	O‘zlik olmoshi	*óz*
	+Pron_Int	So‘roq olmoshlari	*kim, ne, qanday, neshe*
	+Pron_Dem	Ko‘rsatish olmoshlari	*bul, sol, usı*
	+Pron_Det	Belgilash olmoshlari	*hárkim, hárbir*
	+Pron_Coll	Jamlovchi olmoshlar	*barlıq, hámme*
	+Pron_Neg	Inkor olmoshlari	*hesh kim, hesh nárse*
	+Pron_Indef	Belgisizlik olmoshlari	*állekim, állebir*
Son (Numeral)	+Num_Infl	Sonlar	*bir, on, júz, million*
Taqlid so‘zlar (Onom.)	+Onom_Infl	Seske eliklewishler	*tars, dúrs, tırs*
	+Ideo_Infl	Tasvirga taqlidlar	*jalt, jarq, selk*
Bog‘lovchi (Conj.)	+Conj_Coord	Biriktiruvchi bog‘lovchilar	*hám, da, de, menen*
	+Conj_Adverbs	Zidlovchi bog‘lovchilar	*biraq, lekin*
	+Conj_Disj	Ko‘chma bog‘lovchilar	*ya, yaki, yamasa*

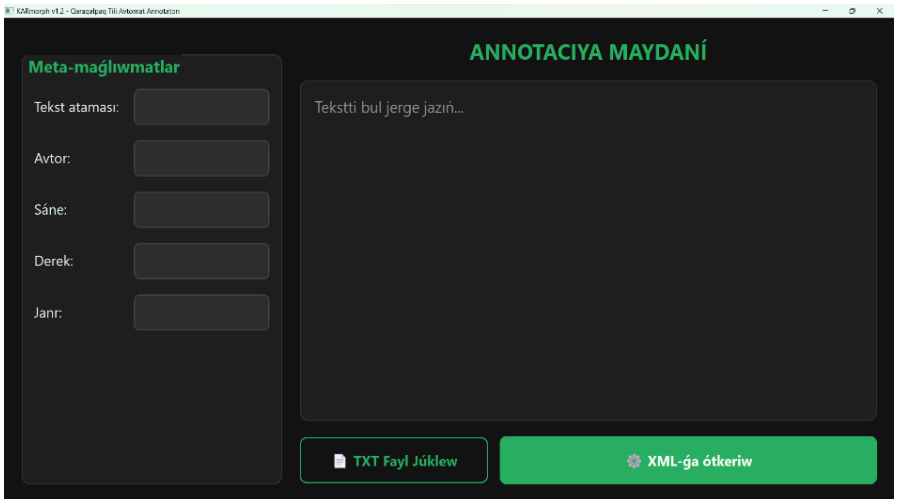
	+Conj_Cond	Shart bog‘lovchilar	*eger, eger de*
	+Conj_Causal	Sabab bog‘lovchilar	*óytkeni, sebebi*
Undov (Interj.)	+Intj_Cont	Undovlar	*pah, pay, yasha, átteń*
Yuklama (Particle)	+Ptc_Ques	So‘roq birligi	*-ma, -me, -ba, -be*
	+Ptc_Lim	Chegaralovchi tub son	*tek, ǵana*
	+Ptc_Emph	Kuchaytirish birligi	*eń, júdá, oǵada*
	+Ptc_Mod	Modal tub son	*aw, ay, shı/shi*

Morfoannotatorning asosiy vazifalari

Qoraqalpoq tili uchun yaratilayotgan morfoannotator dasturi quyidagi tarzda yaratiladi, eng avval foma muhitida yaratilgan KARmorph.foma morfoanalizatorining binar faylidan foydalanamiz, bunda python dasturlash tili bilan foma muhitini integratsiya qilamiz. Python dasturi KARmorph.bin binar fayliga tayanib, undagi ma’lumotlardan foydalangan holda berilgan matnni tahlil qilib beradi. Python skriptiga qo‘yiladigan talablar ham yo‘q emas. Ularning qatoriga dastur interfeysining bo‘lishi, interfeys o‘z ichiga ma’lumotlarni yuklash qismlarini olishi kerak, ularning qatoriga korpus bazasini yaratishda eng muhim ma’lumotlardan bo‘lgan meta-ma’lumotlarni kiritish oynalari bo‘ladi: janr <source><source/>, muallif <author> </author>, yil va sana <date></date>, toifa, matn nomi: <title> </title> va h.k. Interfeys tarkibida matnlarni yuklash, matnni tahlil qilish va xml ma’lumotlar bazasiga aylantiruvchi funksiyalar ham bo‘lishi kerak.

XML ma’lumotlar bazasini yaratishda korpus lingvistikasiga an’anaga aylangan iyerarxiyadan foydalanish maqsadga muvofiq. Masalan, matn eng katta birlik bo‘lsa-da, biz uning ichidagi katta birliklarga ham e’tibor qaratishimiz kerak. Eng avvalo matnda beriladigan meta ma’lumotlar qatorini shakllantirishda matn

nomi, muallif, matnga havola, sana, janr, kategoriya kabi ma’lumotlarni e’tibordan chetda qoldirmadik.



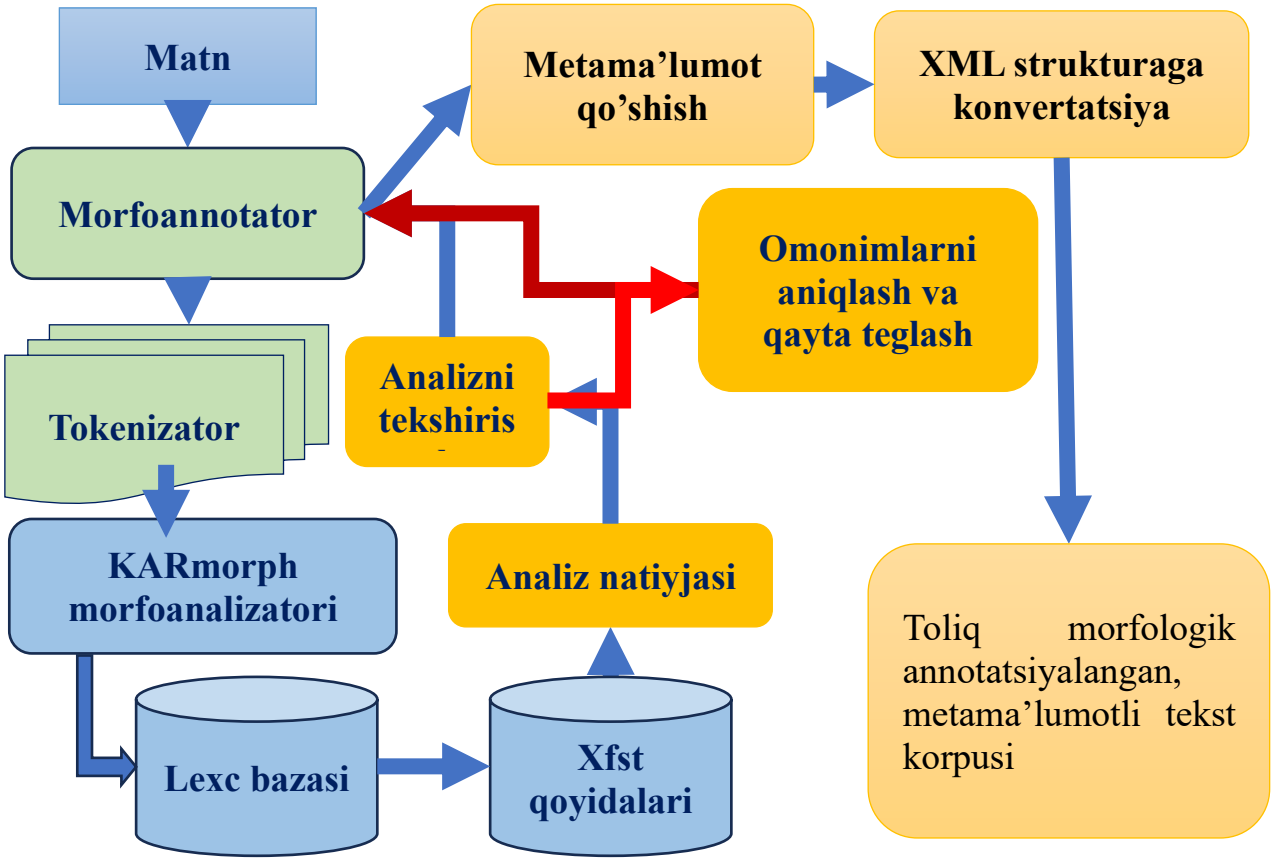
2-rasm. Avtomat annotatsiyalovchi dasturning interfeysi

Matn tarkibidagi elementlar: abzas paragraph 1 , gap <phrase>, so‘z birikmasi, so‘z <words>, morfema va uning turlari <word>, <item></item>. Bularning barchasini biz xalqaro standartlar asosida quyidagi tuzilmada berib o‘tdik:

```
<?xml version="1.0" ?>
<lingPaper automaticallywrapinterlinears="yes">
  <frontMatter>
    <title> </title>
    <author> </author>
    <date></date>
    <source><source/>
  </frontMatter>
  <section1 id="s1">
    <secTitle>Interlinear text</secTitle>
    <p>paragraph 1</p>
    <example num="xf731b03c">
      <interlinear>
        <phrase>
          <words>
            <iword>
                        <item  type="txt"  lang="kaa-Latn-
baseline"></item>
                    <item  type="gls"  lang="kaa-wordGloss">
</item>
```

```
                    <item type="pos" lang="kaa-wordCategory">
</item>
                </iword>
              </words>
            </phrase>
          </interlinear>
        </example>
      </section1>
    </lingPaper>
```

Morfoannotator bosqichlari quyidagi tartibda amalga oshiriladi:



1-chizma. KARmorph morfoannotatorining ko‘p bosqichli integratsiyalangan ish sxemasi

Bunda omonimiya masalasini yechish uchun “Omonimlarni aniqlash va qayta teglash” moduli asosiy filtr vazifasini bajaradi.

Xulosa. So‘zimizning yakuni sifatida qoraqalpoq tili uchun bunday annotatorlarning yaratilishi kelajakda katta hajmli matnlarni mashina-odam

ishtirokida annotatsiyalash va bu orqali katta hajmli korpuslarni xalqaro standartlar asosida ishlab chiqish imkoniyatini beradi. Albatta, biz faqat morfologik annotatsiya bilan cheklanib qolmoqchi emasmiz. Annotatorimizning semantik, sintaktik jihatlarini ham yaratib, uni takomillashtirib chiqishni rejalashtirganmiz.

Foydalanilgan adabiyotlar ro‘yxati

1. Beesley K. R., Karttunen L. Finite State Morphology. CSLI Publications, 2003. https://web.stanford.edu/~laurik/fsmbook/home.html

2. Jurafsky D., Martin J. H. Speech and Language Processing (3rd ed. draft). Stanford University, 2024. https://web.stanford.edu/~jurafsky/slp3/

3. Hulden M. Foma: a finite-state compiler and library. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2009. Pp. 29–32. https://aclanthology.org/E09-2008.pdf

4. Çöltekin Ç. A Freely Available Morphological Analyzer for Turkish. Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC), 2010. Pp. 820–827.

5. Britto H., Galves C., Ribeiro I., Augusto M., Scher A. P. Morphological annotation system for automated tagging of electronic textual corpora: From English to Romance languages. State University of Campinas, 1998. https://www.tycho.iel.unicamp.br/wiki/arquivos/2/22/BRITTO_Hetal-Fase1a.pdf

6. Durst P., Szabó M. K., Vincze V., Zsibrita J. Using Automatic Morphological Tools to Process Data from a Learner Corpus of Hungarian. Apples – Journal of Applied Language Studies, 8(3), 2014. Pp. 39–54. https://apples.journal.fi/article/view/97871/55884

7. Otemisov A. Jahón tilshunosligida morfologik analizator yaratish va tadqiq qilish tarixidan, in Proc. Int. Sci.-Theor. Conf. Problems of Research and Education of the Uzbek Language, Tashkent, 2022. Pp. 297–301.

8. Otemisov A., Sharbaev J. Morfoanalizator bosqichi uchun qoraqalpoq tilidagi fe’l so‘z turkumini formallashtirish moduli, in Proc. Int. Sci.-Pract. Conf. Contemporary Technologies of Computational Linguistics (CTCL.2024), vol. 2, Apr. 2024. [Online]. Available: https://www.myscience.uz/index.php/linguistics

9. Çöltekin Ç. TRmorph: A free morphological analyzer for Turkish, GitHub repository. [Online]. Available: https://github.com/coltekin/TRmorph

10. Israilova N. A., Bakasova P. S. Morphological analyzer of the Kyrgyz language, in Proc. 5th Int. Conf. on Computer Processing of Turkic Languages «TurkLang 2017», Vol. 2, Kazan, 2017, pp. 100–116.

11. Gilmullin R. A., Ganeev B. T., Suleymanov M. Z. Transition of the Tatar Morphological Analyzer to the HFST Platform, in Proceedings of the 2018 International Conference on Turkic Languages and Information Technologies (TurkLang), Kazan, Russia, 2018, pp. 114–122.