QORAQALPOQ TILINI MODELLASHTIRISHNING NAZARIY ASOSLARI VA MATEMATIK YONDASHUVLARI

Allaberdiyeva Durdana Guranmuratovna
Raqamli texnologiyalar va sun’iy intellektni rivojlantirish
ilmiy-tadqiqot instituti doktoranti
Email: durdonallaberdiyeva39@gmail.com

*DOI: **https://doi.org/10.5281/zenodo.20009604***

***Annotatsiya**: Ushbu maqolada qoraqalpoq tilini modellashtirishning nazariy asoslari, lingvistik va matematik yondashuvlari kompleks tarzda tahlil qilinadi. Xususan, korpus lingvistikasi, ehtimollik nazariyasiga asoslangan til modellari, n-gram yondashuvi va zamonaviy neyron modellar ko‘rib chiqiladi. Agglyutinativ tillarga xos morfologik murakkabliklar, ma’lumotlar kamligi va siyraklik muammolari ilmiy asosda tahlil qilinadi. Tadqiqot natijalari qoraqalpoq tilining raqamli modellashtirilishi va tabiiy tilni qayta ishlash tizimlarini ishlab chiqishda muhim ahamiyat kasb etadi.*

***Kalit so‘zlar:** qoraqalpoq tili, til modeli, n-gram, ehtimollik modeli, korpus lingvistikasi, NLP, agglutinatsiya.*

***Аннотация**: В статье рассматриваются теоретические основы моделирования каракалпакского языка с использованием лингвистических и математических методов. Особое внимание уделяется корпусной лингвистике, вероятностным моделям, n-граммам и нейронным сетям. Анализируются проблемы агглютинативных языков, включая морфологическую сложность и разреженность данных. Результаты исследования могут быть использованы при разработке систем обработки естественного языка.*

***Ключевые слова:** каракалпакский язык, языковая модель, n-грамма, вероятностная модель, корпусная лингвистика, NLP.*

***Abstract:** This paper investigates the theoretical foundations of modeling the Karakalpak language using linguistic and mathematical approaches. It focuses on corpus linguistics, probabilistic models, n-gram models, and neural network-based approaches. Special attention is given to the challenges of agglutinative languages, including morphological complexity and data sparsity. The findings are important for developing natural language processing systems and digital language resources.*

Keywords: *Karakalpak language, language model, n-gram, probabilistic model, corpus linguistics, NLP.*

Kirish. Tilni modellashtirish zamonaviy kompyuter lingvistikasi va tabiiy tilni qayta ishlash (NLP) sohasining asosiy yo‘nalishlaridan biri hisoblanadi. Til modeli tabiiy til birliklarini matematik shaklda ifodalash va ularning ehtimollik taqsimotini aniqlash imkonini beradi. Umumiy holda til modeli so‘zlar ketma-ketligining ehtimolligini baholash orqali ifodalanadi:

$P(w_1, w_2, ..., w_n)$.

Bu yerda w_1 dan w_n gacha bo‘lgan elementlar matndagi so‘zlardir. Ushbu ehtimollikni aniqlash orqali matnni avtomatik tahlil qilish, generatsiya qilish va tushunish mumkin bo‘ladi. Qoraqalpoq tili turkiy tillar oilasiga mansub bo‘lib, agglutinativ tuzilishga ega. Agglutinatsiya xususiyati so‘zlarga ketma-ket qo‘shimchalar qo‘shilishi orqali yangi grammatik shakllar hosil bo‘lishini anglatadi. Natijada bitta ildizdan juda ko‘p shakllar yuzaga keladi, bu esa modellashtirish jarayonida lug‘at hajmining keskin ortishiga olib keladi. Shu sababli qoraqalpoq tilini modellashtirishda ehtimolliklarni aniqlash, siyraklik muammosini hal qilish va samarali algoritmlar yaratish dolzarb ilmiy masalalardan biridir. Ushbu maqolaning maqsadi qoraqalpoq tilini modellashtirishning nazariy asoslarini tizimli ravishda yoritish, matematik formulalar asosida tahlil qilish va mavjud muammolarni ko‘rsatishdan iborat.

Adabiyotlar sharhi. Til modellashtirish masalalari dastlab statistik yondashuvlar asosida shakllangan bo‘lib, keyinchalik neyron tarmoqlar bilan boyitildi. Jurafsky va Martin (2009) tomonidan til modellarining ehtimollik asoslari batafsil yoritilgan bo‘lib, ular tilni ehtimollik taqsimoti sifatida qarashni taklif etadi [1]. Statistik modelga ko‘ra, gapdagi har bir so‘z oldingi so‘zlarga bog‘liq holda yuzaga keladi va bu quyidagicha ifodalanadi:

$P(w_1, w_2, ..., w_n) = \prod P(w_i \mid w_1, ..., w_{i-1})$.

Ammo ushbu formulani amaliy hisoblash juda murakkab bo‘lgani sababli Markov taxmini qo‘llaniladi. Manning va Schütze (1999) o‘z tadqiqotlarida n-gram modellari til modellashtirishning eng muhim vositalaridan biri ekanligini ko‘rsatadi [2]. N-gram modeli kontekstni cheklangan miqdordagi oldingi so‘zlar orqali ifodalaydi. Masalan, bigram modeli $P(w_i \mid w_{i-1})$ ko‘rinishida, trigram esa $P(w_i \mid w_{i-2}, w_{i-1})$ ko‘rinishida ifodalanadi. Biroq ushbu modellarning asosiy muammosi siyraklik bo‘lib, kam uchraydigan kombinatsiyalar uchun ehtimollikni aniqlash qiyinlashadi. Mikolov (2010) tomonidan taklif etilgan neyron til modellari ushbu muammoni qisman hal qilib, yashirin qatlamlar

orqali kontekstni chuqurroq o‘rganishga imkon beradi [3]. Turkiy tillar, jumladan qoraqalpoq tili uchun esa morfologik murakkablik alohida muammo hisoblanadi. Abduraxmonova (2019) korpus lingvistikasi asosida turkiy tillarni modellashtirishda morfologik segmentatsiya va analizning ahamiyatini ta’kidlaydi [4]. Shuningdek, Zipf qonuni (1949) tabiiy tillarda so‘z chastotalarining notekis taqsimlanishini ko‘rsatadi va bu til modellarini qurishda muhim nazariy asos bo‘lib xizmat qiladi [6].

Metodologiya. Qoraqalpoq tilini modellashtirish bir necha bosqichlardan iborat. Birinchi bosqich — korpus yaratish bo‘lib, u matnlar to‘plami sifatida $D = \{S_1, S_2, ..., S_m\}$ ko‘rinishida ifodalanadi. Bu yerda har bir S_i alohida gapni bildiradi. Ikkinchi bosqich — tokenizatsiya jarayoni bo‘lib, matn so‘zlarga ajratiladi: $S = \{w_1, w_2, ..., w_n\}$. Uchinchi bosqichda ehtimolliklar baholanadi. Maksimal ehtimollik bahosi (MLE) quyidagi formula bilan aniqlanadi: $P(w_i \mid w_{i-1}) = Count(w_{i-1}, w_i) / Count(w_{i-1})$. Biroq nol ehtimollik muammosi yuzaga kelganda silliqlash usullari qo‘llaniladi. Masalan, Laplace smoothing: $P(w_i \mid w_{i-1}) = (Count(w_{i-1}, w_i) + 1) / (Count(w_{i-1}) + V)$, bu yerda V lug‘at hajmi hisoblanadi. Agglutinativ tillarda subword modeling muhim ahamiyatga ega bo‘lib, so‘zlar kichik birliklarga ajratiladi: $Word = Subword_1 + Subword_2 + ... + Subword_k$. Bu usul siyraklik muammosini kamaytirishga yordam beradi.

Natijalar. Tahlillar shuni ko‘rsatadiki, qoraqalpoq tilida morfologik variantlar soni juda katta bo‘lib, bu lug‘at hajmining ortishiga olib keladi. Natijada n-gram modellari samaradorligi pasayadi va ehtimolliklarni aniq baholash qiyinlashadi. So‘z chastotalari Zipf qonuniga bo‘ysunadi, ya’ni $f(r) \approx 1/r$. Bu esa kam uchraydigan so‘zlar soni juda ko‘pligini anglatadi. Korpus hajmi kichik bo‘lgan holatda modelning aniqligi sezilarli darajada kamayadi. Shu bilan birga, subword yondashuvlar va silliqlash usullari model natijalarini yaxshilashga xizmat qiladi.

Muhokama. Qoraqalpoq tilini modellashtirish jarayoni zamonaviy kompyuter lingvistikasi va tabiiy tilni qayta ishlash (Natural Language Processing) yo‘nalishidagi muhim masalalardan biri hisoblanadi. Ushbu tadqiqot doirasida tilning fonetik, morfologik va sintaktik qatlamlarini matematik va statistik modellar orqali ifodalash imkoniyatlari ko‘rib chiqildi. Olingan natijalar shuni ko‘rsatadiki, qoraqalpoq tili agglutinativ xususiyatga ega bo‘lgani sababli, uni modellashtirishda oddiy chiziqli yondashuvlar yetarli emas, balki murakkab ehtimollik modellari va neyron tarmoqlarga asoslangan metodlar talab etiladi.

Tadqiqot davomida korpus asosida ishlashning ahamiyati alohida ta’kidlandi. Ayniqsa, qoraqalpoq tilida katta hajmli va belgilangan (annotatsiyalangan) matnlar bazasining yetishmasligi modellashtirish sifatiga sezilarli ta’sir ko‘rsatadi. Shu sababli, til korpusini yaratish va uni doimiy ravishda boyitib borish model aniqligini oshirishda hal qiluvchi omil bo‘lib xizmat qiladi. Bu jihat korpus lingvistikasi doirasida olib borilayotgan tadqiqotlar bilan hamohangdir.

Modellashtirishda qo‘llanilgan yondashuvlar orasida n-gram modellari, yashirin Markov modellari hamda chuqur o‘rganish algoritmlariga asoslangan usullar taqqoslandi. Natijalar shuni ko‘rsatadiki, an’anaviy statistik modellarga nisbatan sun’iy neyron tarmoqlar asosidagi modellar yuqori aniqlik va moslashuvchanlikni ta’minlaydi. Biroq, bunday modellar katta hajmdagi o‘quv ma’lumotlarini va hisoblash resurslarini talab qiladi.

Shuningdek, morfologik boylik darajasi yuqori bo‘lgan qoraqalpoq tilida so‘z shakllarining xilma-xilligi modellashtirish jarayonini murakkablashtiradi. Bu muammoni hal etishda subword (bo‘g‘in yoki affiks asosida segmentatsiya) texnikalari samarali yechim sifatida ko‘rib chiqildi. Shu bilan birga, tilning sintaktik tuzilishini chuqur o‘rganish va uni modellashtirish uchun dependensiya grammatikasi va transformatsion yondashuvlardan foydalanish istiqbolli natijalar berishi mumkin. Umuman olganda, tadqiqot natijalari qoraqalpoq tilini modellashtirishda kompleks yondashuv zarurligini ko‘rsatadi. Ya’ni, lingvistik bilimlar, matematik modellar va zamonaviy dasturiy texnologiyalar integratsiyasi yuqori samaradorlikka erishish imkonini beradi.

Xulosa. Mazkur tadqiqotda qoraqalpoq tilini modellashtirishning nazariy va amaliy asoslari keng qamrovda tahlil qilindi. Tadqiqot natijalari shuni ko‘rsatdiki, tilning strukturaviy xususiyatlarini chuqur o‘rganish va ularni matematik modellar orqali ifodalash zamonaviy lingvistik texnologiyalarni rivojlantirishda muhim ahamiyatga ega. Qoraqalpoq tilining agglyutinativ tabiati, morfologik boyligi va sintaktik moslashuvchanligi uni modellashtirish jarayonini murakkablashtirsa-da, zamonaviy sun’iy intellekt va mashinaviy o‘rganish usullari bu muammoni samarali hal etish imkonini bermoqda. Ayniqsa, neyron tarmoqlar va chuqur o‘rganish algoritmlarining qo‘llanilishi til modelining aniqligini sezilarli darajada oshiradi. Tadqiqot davomida aniqlangan asosiy muammolardan biri — sifatli va keng qamrovli til korpusining yetishmasligidir. Shu sababli, kelgusida qoraqalpoq tilining elektron korpusini yaratish, uni standartlashtirish va ochiq foydalanishga taqdim etish ustuvor vazifa sifatida qaralishi lozim. Bu nafaqat tilni modellashtirish, balki avtomatik tarjima, nutqni tanish va matnni tahlil qilish kabi sohalarda ham muhim natijalarga olib keladi.

Xulosa qilib aytganda, qoraqalpoq tilini modellashtirish istiqbolli ilmiy yo‘nalish bo‘lib, u tilning raqamli muhitdagi rivojlanishini ta’minlashga xizmat qiladi. Ushbu yo‘nalishda olib boriladigan kelgusidagi tadqiqotlar yanada mukammal modellarning yaratilishiga va qoraqalpoq tilining global axborot makonida o‘z o‘rnini mustahkamlashiga zamin yaratadi.

Foydalanilgan adabiyotlar.

1. Jurafsky, D., Martin, J. H. Speech and Language Processing. Pearson, 2009.
2. Manning, C. D., Schütze, H. Foundations of Statistical Natural Language Processing. MIT Press, 1999.
3. Mikolov, T. Neural Network Based Language Model. 2010.
4. Abduraxmonova, N. Korpus lingvistikasi bo‘yicha ilmiy tadqiqotlar. 2019.
5. Koehn, P. Statistical Machine Translation. Cambridge University Press, 2010.
6. Zipf, G. K. Human Behavior and the Principle of Least Effort. 1949.