

QORAQALPOQ TILI SO‘ZLARINI LEMMALASH UCHUN LINGVISTIK MANBA YARATISH

Yusupov Firnafas

Texnika fanlari nomzodi, dotsent

firnafas@mail.ru

TATU Urganch filiali

Sharipov Maqsud Siddiqovich

Texnika fanlari nomzodi, dotsent

maqsbek72@gmail.com

Urganch davlat universiteti

Urazbayeva Nilufar Oybek qizi

urazbayevanilufar724@gmail.com

UrDU magistranti

Annotatsiya. Mazkur maqolada qoraqalpoq tilidagi so‘zlarni lemmalash uchun zarur bo‘lgan lingvistik manbani yaratish masalasi ko‘rib chiqiladi. Unda qoraqalpoq tilining leksik xususiyatlari, affiksial tizimi, shuningdek, so‘zlarning lemmalash jarayonidagi o‘rni tahlil qilinadi. Leksik tuzilmalarni aniqlash, lug‘aviy bazani shakllantirish lemmalashning muhim bosqichlari sifatida ko‘rsatiladi. Mualliflar qoraqalpoq tiliga oid leksik va grammatik axborotlarni tizimlashtirish orqali, tabiiy tilni qayta ishlash, mashina tarjimai va sun‘iy intellekt sohalarida foydalanish mumkin bo‘lgan lingvistik manbani ishlab chiqish zarurligini asoslab beradilar. Tadqiqot natijalari qoraqalpoq tilining raqamli muhitda rivojlanishiga hissa qo‘shadi.

Annotation. This article discusses the issue of creating a linguistic resource necessary for the lemmatization of words in the Karakalpak language. It analyzes the lexical features and affixal system of the Karakalpak language, as well as the role of words in the lemmatization process. Identifying lexical structures and building a lexical database are presented as important stages of lemmatization. The authors justify the need to develop a linguistic resource that can be used in natural language processing, machine translation, and artificial intelligence by systematizing lexical and grammatical information related to the Karakalpak language. The research results will contribute to the development of the Karakalpak language in the digital environment.

Аннотация. В данной статье рассматривается вопрос создания лингвистического ресурса, необходимого для лемматизации слов каракалпакского языка. Анализируются лексические особенности, аффиксальная система каракалпакского языка, а также роль слов в процессе

лемматизации. Определение лексических структур и формирование лексической базы представлены как важные этапы лемматизации. Авторы обосновывают необходимость разработки лингвистического ресурса, который можно использовать в области обработки естественного языка, машинного перевода и искусственного интеллекта, путем систематизации лексической и грамматической информации, относящейся к каракалпакскому языку. Результаты исследования внесут вклад в развитие каракалпакского языка в цифровой среде.

Kalit so‘zlar. *Qoraqalpoq tili, korpus, korpus tahlili, lingvistik manba, lemma, lemmalash, algoritm.*

Hozirgi kunda qoraqalpoq tilidagi lemmalash jarayoniga oid ilmiy tadqiqotlar va manbalar yetishmayapti. Shu sababli, qoraqalpoq tilidagi so‘zlarni lemmalash uchun lingvistik manba yaratish zarurati tug‘iladi. Ushbu maqolada qoraqalpoq tili so‘zlarini lemmalash uchun kerakli lingvistik manbaning yaratilishi jarayoni, uning ahamiyati, Tabiiy tilni qayta ishlash (ing. Natural Language Processing – NLP) muammolarini yechishga yo‘naltirilgan resurslarga e‘tibor qaratish va kutilayotgan natijalari haqida batafsil ma‘lumot beriladi. Lemmalash so‘zlarning asosiy shaklini aniqlash va ularga ishlov berish jarayonini ifodalaydi, bu esa to‘liq lingvistik ma‘lumotlar bazasini talab etadi. Mazkur ma‘lumotlar bazalari korpus deb nomlanadi va u yozma matnlarning istalgan to‘plamidan iborat bo‘lishi mumkin [1]. Lingvistik tahlil dasturlari ushbu korpuslardan tezkor qidiruv va tahlilni amalga oshirish orqali samarali natijalarga erishadi. Maqsadimiz qoraqalpoq tilining raqamli muhitda rivojlanishini ta‘minlash, uning lingvistik xususiyatlarini o‘rganish va xalqaro tilshunoslik tadqiqotlariga hissa qo‘shishdir.

Qoraqalpoq tili – turkiy tillar qatoriga kiradi. 1 mlnga yaqin insonlar ushbu tilda so‘zlashadilar. Qoraqalpog‘iston Respublikasida istiqomat qiluvchi qoraqalpoq xalqining ona tilidir. Qoraqalpoq tili – O‘zbekiston Respublikasining tarkibidagi Qoraqalpog‘iston Respublikasining rasmiy tilidir. Bu to‘g‘risida maxsus qonun chiqarilgan. Qonunda qoraqalpoq tili Qoraqalpog‘istonning barcha hududida davlat tili sifatida foydalanishga imkon berilgan. Qoraqalpoq tili - turkiy tillarning qipchoq guruhiga mansub tillardan; qozoq va no‘g‘ay tillari bilan birgalikda qipchoq tillarining qipchoq-no‘g‘ay guruhchasini tashkil etadi. Qoraqalpog‘iston Respublikasining davlat tili (o‘zbek tili bilan birga). Asosan, Qoraqalpog‘istonda, shuningdek, Xorazm, Navoiy, Buxoro viloyatlari hamda Qozog‘iston va Turkmanistonning unga qo‘shni hududlarida, Rossiya Federatsiyasi va Afg‘onistonda tarqalgan. Qoraqalpoq tilida so‘zla-shuvchilarning umumiy soni 883 ming kishidan iborat (o‘tgan asrning 90-yillar o‘rtalari). Qoraqalpoq tili, asosan, 2 ta: shimoli-sharqiy va janubi-g‘arbiy lahjalarga bo‘linadi; bu lah-jalar fonetik jihatdan o‘zaro farqlanadi. Adabiy Qoraqalpoq tili 20-asrning 1-yarmida shimoli-

sharqiy lahja asosida shakllangan. Qoraqalpoq tili boshqa turkiy tillarning aksariyati kabi tipologik xususiyatlari bilan ajralib turadi [2, 3, 4].

Qoraqalpoq tilining o'ziga xos belgilari: unlilar ohangdoshligi - singarmonizm mavjud, Atlarshmz (otlarimiz). kunler - (kunlar); umumturkiy ch undoshi sh bilan; sh esa s undoshi bilan almashadi, qash (qoch), tas (tosh), bas (bosh). Ayrim so'zlarda g' undoshi o'rnida v; g undoshi o'rnida y ishlatiladi, tav (tog'), tiy (teg) va boshqa Qoraqalpoq tilida, boshqa ko'pchilik turkny tillarda bo'lganidek, umumturkiy so'zlardan tashqari, arab, eroniy, rus tillaridan o'zlashgan so'zlar ham anchagina [3, 4].

Hozirgi kunda, tillarning kompyuter yordamida ishlashini o'rganuvchi sohalarida, lemmalash muhim ahamiyatga ega. Bu maqolada, qoraqalpoq tilida lemmalash algoritmini yaratish uchun zarur bo'lgan lingvistik manbalarni ishlab chiqish masalasi ko'rib chiqiladi. Qoraqalpoq tili, turkiy tillar oilasiga mansub bo'lib, o'zining leksik, sintaktik va morfologik xususiyatlari bilan ajralib turadi. Bu tilda ko'plab o'zgartirishlar va afikslar mavjud, ular so'zning ma'no va grammatik shaklini o'zgartiradi. Qoraqalpoq tilining leksikologiyasini o'rganish, lemmalash algoritmini yaratishda asosiy bosqichdir. Qoraqalpoq tilida fe'llar ko'p hollarda zamon va shaxs affikslariga ega. Shuningdek, nomlar uchun ko'plik, egalik, to'g'ridan-to'g'ri va bilvosita qo'shish affikslaridan foydalaniladi. So'zning har bir shakli o'zining lemmataga keltirilishini talab qiladi, shuning uchun qoraqalpoq tilining morfologik xususiyatlarini batafsil tushunish, algoritmnining samaradorligini oshiradi [4, 5].

Lemmalashish (yoki lemmatization) – bu so'zning morfologik shakllarini uning lemmaga (ya'ni, so'zning ildiz shakliga) keltirish jarayonidir. Masalan, turk tilida “yazacak”, “yazdi”, “yaziyor” kabi shakllar bir xil lemmaga, ya'ni "yaz" ga keltiriladi. Qoraqalpoq tilida ham so'zlar turli morfologik shakllarda bo'lishi mumkin, shuning uchun lemmalash jarayoni ushbu tilning to'g'ri tahlil qilinishi uchun zarurdir.

- Inglizcha. “ran”, “runs”, “running” -> “run”
- O'zbekcha. “yozdi”, “yozadi”, “yozyapti”, “yozgan” -> “yozmoq”
- Qoraqolpoqcha. “yazdi”, “yazaman”, “yazish” – “yazmaq”
- Turkcha. “koşacak”, “koşuyor”, “koşdı” – “koş”

Lemmalash algoritmini yaratish uchun tilga oid katta hajmdagi lingvistik manbalarni tayyorlash zarur. Bu manbalar quyidagi turlardan iborat bo'lishi mumkin:

Qoraqalpoq tilidagi so'zlarning leksik tuzilishini o'z ichiga olgan lug'atlar. Ular so'zlarning lemmalash uchun asosiy manba bo'lib xizmat qiladi. Buning uchun

Qoraqalpoq tilining 7 tomlik izohli lugʻatini misol keltirishimiz va undan foydalanishimiz mumkin. Ushbu izohli lugʻatda har bir harf uchun alohida taʼrif berilib oʻtilgan va har bir soʻzning ikki va undan ortiq shakllari ham koʻrsatilgan. Bu lugʻatdagi har bir harfdagi soʻzlarni alohida fayl shaklida yigʻib chiqishni talab qiladi.

Qoraqalpoq tili uchun lemmalashni amalga oshirishda lingvistik manba yaratishning muhim ahamiyati bor. Bu manbalar tilni avtomatik qayta ishlashda, mashina oʻrganish va sunʼiy intellekt tizimlarini yaratishda yordam beradi. Lemmalashish uchun quyidagi manbalarni yaratish mumkin:

Lingvistik lugʻat: Qoraqalpoq tilida lingvistik lugʻat yaratish uchun - soʻzlarning ildiz shakllarini har xil qoʻshimchalardan holi tarzda koʻrsatib berish. Bu lugʻat tilning leksik va lugʻaviy shakli haqida maʼlumot beradi. Misol uchun, “yozgansizlar” soʻzining qoʻshimchasiz shakli “yoz”, “kitobdagi” soʻzining lemmasi esa “kitob” ekanligi - bularning barchasi lugʻatda oʻz aksini topadi.

Leksik analizator: Lemmalashish jarayonini avtomatlashtirish uchun leksik analizator yaratish zarur. Bu tizim, kiruvchi matndagi har bir soʻzni analiz qilib, uni lemmaga keltiradi. Masalan, "kitoblar" soʻzi "kitob" lemmaga, "yazaman" soʻzi esa "yaz-" ildiziga keltiriladi. Bunday analizator dasturlari Qoraqalpoq tilining xususiyatlarini hisobga olgan holda ishlab chiqilishi kerak.

Sentiment tahlilini amalga oshirish: Lemmalashish orqali tilda ifodalangan xabar yoki fikrlarni oʻrganish imkoniyati oshadi. Masalan, “tezroq” va “tezlik” soʻzlarining lemmalashidan keyin, bir xil semantik maʼno (tezlik) oʻrganilishi mumkin. Bu esa sentiment tahlili kabi sohalarda foydali boʻladi.

Tegishli korpuslar yaratish: Qoraqalpoq tilini lemmalash uchun maxsus til korpuslari kerak boʻladi. Bunday korpuslar tilning tabiiy materiallarini oʻz ichiga oladi va soʻzning kontekstual oʻzgarishlarini tahlil qilishda yordam beradi. Korpusdagi soʻzlar, iboralar va ularning grammatika qoidalari lemmalashish jarayonini toʻgʻri amalga oshirishda asos boʻlib xizmat qiladi.

Xulosa qilib shuni aytish lozimki, lingvistik muammolarni hal qilishda, tarjima dasturlari ishlab chiqishda lingvistik manbaning oʻrni katta. Ushbu maqolada aynan qoraqalpoq tili uchun lingvistik manba yaratishning nazariy va amaliy maʼlumotlari berib oʻtildi. Qoraqalpoq tili uchun lemmalash algoritmini yaratish, turkiy tillarning morfologiyasini va lingvistik xususiyatlarini chuqur tushunishni talab qiladi. Tabiiy tilni qayta ishlashning samaradorligini oshirish uchun qoraqalpoq tiliga oid toʻliq va toʻgʻri lingvistik manbalar yaratish zarur. Shuningdek, kompyuter lingvistik yondoshuvlar, masalan, mashina oʻrganish va statistik metodlar yordamida, lemmatizatsiya jarayonini yanada takomillashtirish mumkin.



Bu tadqiqot, qoraqalpoq tilining tabiiy tilni qayta ishlash sohasida muvaffaqiyatli rivojlanishiga hissa qo'shishi mumkin. Lemmalash algoritmining samaradorligi, tilning tabiiy shakllaridan foydalanishda va avtomatik tarjimada katta yordam beradi. Kelajakda ushbu manbani mukammallashtirib, uni turli xil janrdagi matnlarni atroflicha tahlil qilish uchun asosiy manbaga aylantirmoqchimiz.

Foydalanilgan adabiyotlar:

1. Baskakov N. A. Karakalpakskiy yaznk, t. 1 – 2, M. – L., 1951-52;
2. Berdimuratov Ye., Hazirgi zaman qaraqalpoq tilinip lskologiyasm, Nokis, 1968;
3. Baskakov N. A., Vvedeniye v izucheniye tyurskix yazmkov, 2-izd., M., 1969;
4. Berdimuratov Ye. Umarov A. Bekimbetov P., Karakalpakskiy yazj (fonetika, morfologiya, sintaksis), Nukis. 1983.
5. Birgit N. Schlyter. Language Policies in Present-Day Central Asia(ingl.)// International Journal on Multicultural Societies. -UNESCO, 2001. -Iss. Vol. 3, No. 2. -P. 130. -ISSN 1817-4574. Архивировано 10 април 2018 года.
6. Қарақалпак тилинин түсиндирме сөзлиги.
7. О'zbekiston milliy ensiklopediyasi (2000-2005).