

## TÁBIYIY HÁM TEXNIKA LIQ ILIMLER

**Fizika. Matematika. Texnika. Informatika.**

### QARAQALPAQ TILINI LINGVISTIKALIQ KORPUS PLATFORMASIN JARATI W

(Tólepbergen Qayıpbergenovtıń shıǵarmaları misalında)

**G.R.Abdaliev** – *filologiya ilimleriniń kandidatu, docent*

**N.U.Uteuliev** – *fizika-matematika ilimleri doktorı, professor*

**B.K.Kalmuratov** – *áǵa oqıtıwshı*

**A.S.Utesinova** – *magistrant*

*Tashkent informaciyalıq texnologiya universiteti Nókis filiali*

**Tayanch soʻzlar:** qoraqalpoq tili, lingvistik korpus, maʼnaviy meros, morfologik tahlil, sintaktik tahlil, semantik tahlil, lemmatizatsiya, tokenizatsiya, sunʼiy intellekt, frazeologizmlar.

**Ключевые слова:** каракалпакский язык, лингвистический корпус, духовное наследие, морфологический анализ, синтаксический анализ, семантический анализ, лемматизация, токенизация, искусственный интеллект, фразеологизмы.

**Key words:** karakalpak language, linguistic corpus, cultural heritage, morphological analysis, syntactic analysis, semantic analysis, lemmatization, tokenization, artificial intelligence, phraseology.

Házirgi dáwirde sanlı texnologiyalardıń rawajlanıwı nátiyjesinde jámiyetimizdiń barlıq tarawları sanlastırılmaqta. Usınıń nátiyjesinde zamanagóy joqarı texnologiyalardan paydalanıw múmkinshiligi arta beredi. Ásirese, dúnya júzinde aldınnan tayarlangan jasalma intellekt til modelleriniń nátiyjesinde úlken jetiskenlikler ámelge asırılmaqta. Usılardıń qatarında qaraqalpaq til modellerin jetilistiriw, lingvistikalıq hám de parallel korpus jaratıw máseleleri eń aktual zárúrlik bolıp tabıladı.

**Lingvistikalıq korpus** bul – belgili bir tildin yamasa dialektiń úlken kólemdegi tekstler jıynaǵı. Ol til izertlewshileri, oqıtıwshılar hám til texnologiyaları menen shuǵıllanıwshılar ushın áhmiyetli qural bolıp tabıladı. Korpus arqalı tildin haqıyqıy qollanıwın, sóz jıyılıǵın, grammatikalıq qurılısların hám basqa da lingvistikalıq aspektlerin analizlewge boladı.

Lingvistikalıq korpusıń tiykarǵı ózgeshelikleri tómendegilerdi óz ishine aladı. **Reprezentativlik** ózgeshelikte korpus tildin hár qıylı stillerin, janların hám dáwirlerin qamtıwı kerek. Kólemi jaǵınan, ádette, millionlaǵan sózlerden ibarat boladı. Mashina oqıy alatuǵın formatta kompyuter analizi ushın qolaylı bolıwı kerek. **Annotaciyalıq** ózgeshelikte kóbinese, morfologiyalıq, sintaksislik hám semantikalıq belgiler menen belgilenedi [1].

Tólepbergen Qayıpbergenovtıń shıǵarmaları tiykarında lingvistikalıq korpus dúziw, bul – qaraqalpaq tilin izertlew ushın áhmiyetli basqısh bolıp tabıladı. Bul jazıwshınıń shıǵarmaları qaraqalpaq ádebiy tiliniń bay úlgisi bolıp esaplanadı.

Korpus dúziwdiń tiykarǵı basqıshları:

1. Tekstlerdi jıynaw hám sanlastırıw;
2. Tekstlerdi normalizaciyalaw (orfografiyanı birlestiriw, punktuaciyalıq tártipke salıw);
3. Tokenizatsiya (tekstti sózlerge hám gáplerge ajratıw);
4. Morfologiyalıq hám sintaksislik analiz;
5. Metamaǵlıwmatlardı qosıw (shıǵarmanıń atı, jazılǵan jılı hám t.b.) [2].

Tólepbergen Qayıpbergenovtıń shıǵarmaları boyınsha korpus platformasınıń islew - bul júdá áhmiyetli joybar. Jumıstı baslaw ushın tómendegilerdi ámelge asıramız:

**Materiallardı jıynaw**, yaǵnıy T.Qayıpbergenovtıń barlıq shıǵarmaların jıynaw hám olardı sanlı formatqa ótkeriw.

**Tekstlerdi tayarlaw** arqalı shıǵarmalardı bóleklerge bóliw (misalı, betler, bólimler yamasa abzaclar boyınsha). Hár bir bólekke metamaǵlıwmatlardı qosıw (shıǵarmanıń atı, jazılǵan jılı hám t.b.).

**Lingvistikalıq analizdi ámelge asırıw.** Yaǵnıy tekstlerdi morfologiyalıq hám sintaksislik analizden ótkeriw. Sózlerdi **lemmatizaciyalaw** (sózlerdiń tiykarǵı formasın anıqlaw).

Barlıq maǵlıwmatlar sanlastırılǵannan keyin korpus bazasını jaratıw ámelge asırıladı. Bunda, analizlengen tekstlerdi saqlaw ushın maǵlıwmatlar bazasını tańlaw hám dúziw ámelge asırıladı. Tekstlerdi hám olarǵa baylanıslı **metamaǵlıwmatlardı bazaǵa kirgiziw.** Al, keyingi basqıshta **izlew sistemasın islep shıǵıw** ámelge asırıladı. Yaǵnıy korpusıta izlewdi ámelge asırıw ushın algoritmler islep shıǵıw hám sózler, sóz dizbekleri hám grammatikalıq formalar boyınsha izlew múmkinshiliklerin jaratıw [3].

Soń **platforma interfeysi** jaratıladı. Jazıwshınıń shıǵarmalarınń tekstleri qaraqalpaq tiliniń leksikası, morfologiyası, sintaksisi hám stilistikası jaǵınan tómendegi obyektler boyınsha analizlenedi hám kategoriyalarǵa bólinedi: **Naqıl-maqallar, Frazeologizmler, Metaforalar, Ráwiyatlar, Teńewler, Aforizmler, Epitetler** hám t.b.

Tólepbergen Qayıpbergenovtıń shıǵarmalarındaǵı til hám stillik qurallardı usılay kategoriyalarǵa bóliw korpusıń ilimiy hám ámeliy qunın asıradı. Mine, usı kategoriyalar boyınsha bir neshe itimal jobalar:

**Naqıl-maqallar.** Olardı tematikalıq toparlarǵa bóliw (misalı, turmis haqqında, doslıq haqqında, t.b.). Hár bir naqıl-maqaldıń mánisin túsindiriw. Eger múmkin bolsa, olardıń dáreklerin kórsetiw (xalıq awızeki dóretpesi me yamasa avtorıń ózi dóretken be).

**Frazeologizmler.** Olardı mánisi boyınsha toparlarǵa bóliw. Yaǵnıy hár bir frazeologizmiń sózbe-sóz hám awıspalı mánisin túsindiriw. Olardıń qollanıw jaǵdaylarına misallar keltiriw.

**Metaforalar.** Metaforalardı tematikalıq toparlarǵa bóliw. Bunda olardıń mánisin hám estetikalıq tásirin túsindiriw hám metaforalardıń shıǵarmanıń ulıwma ideyasına qosqan úlesin analizlew.

**Ráwiyatlar.** Ráwiyatlardıń dáreklerin izertlew (eger múmkin bolsa) hám olardıń shıǵarmadaǵı funkciyasın túsindiriw. Ráwiyatlardıń mazmunın hám xızmetin analizlew.

Teñewler. Teñewlerdi qollanılğan obyektleri boyınsha toparlarğa bóliw. Yaǵnıy, olardıń obrazlılıǵın hám tásirsheńligin túsindiriw, sonday-aq, milliy-mádeniy ózgesheliklerin kórsetiw.

Aforizmler. Aforizmlerdi tematikalıq toparlarğa bóliw, olardıń filosofiyalıq hám sociallıq mánisin túsindiriw hám avtorıń dúnyaǵa kózqarasın qalay sáwlelendiretuǵın kórsetiw.

Epitetler. Epitetlerdi semantikalıq toparlarğa bóliw. Bunda olardıń kórkemlik funkciyasın túsindiriw hám avtorıń stiline tiygizgen tásirin analizlew bolıp tabıladi.

Qaraqalpaq ádebiyatınıń kórnekli wákili Tólepbergen Qayıpbergenovtıń shıǵarmaları korpusı, bul – qaraqalpaq tili hám ádebiyatın izertlewde úlken áhmiyetke iye bolǵan jańa hám innovaciyalıq resurs. Bul korpus jazıwshınıń barlıq shıǵarmaların, sonıń ishinde, romanları, gúrınları, publicistikalıq maqalaların óz ishine aladı.

Korpustıń tiykarǵı maqseti – T.Qayıpbergenovtıń shıǵarmaların sanlı formatta saqlaw, olardı izertlewshiler, studentler hám oqıwshılar ushın qolaylı etip usınıw. Bul arqalı qaraqalpaq tili hám ádebiyatın úyreniw procesin jedellestiriw hám jenillestiriw názerde tutılǵan.

Korpustıń tiykarǵı ózgeshelikleri mınalardan ibarat: Toliq qamtıw, yaǵnıy korpus T.Qayıpbergenovtıń barlıq shıǵarmaların óz ishine aladı, bul onıń dóretiwsheńligin tolıq kólemde úyreniwge múmkinshilik beredi. Izlew funkciyası bolsa, korpusıtaǵı sózler, sóz dizbekleri yamasa grammatikalıq konstrukciyalardı izlew múmkinshiligi bar, bul izertlewshilerge kerekli maǵlıwmatlardı tez tabıwǵa járdem beredi. Al, **lingvistikalıq analiz** bóliminde korpustıń morfologiyalıq, sintaksislik hám leksikalıq analiz quralların usınadı, bul T.Qayıpbergenov shıǵarmaları tiliniń ózgesheliklerin tereń úyreniwge imkaniyat beredi. **Statistikalıq maǵlıwmatlar** bólimi arqalı, korpusıtaǵı sózlerdiń kóp ushırasatuǵınlıǵı, kollokaciya hám basqa da statistikalıq maǵlıwmatlardı usınadı, bul jazıwshınıń til qollanıwın analizlewge járdem beredi. Al, **interaktivlik vizualizaciya** arqalı korpus maǵlıwmatların grafikler, diagrammalar hám basqa da vizual qurallar járdeminde kórsetedi, bul maǵlıwmatlardı ańsat túsiniwge imkaniyat beredi.

Korpusı jaratıw arqalı kóplegen nátiyjelerge erisemiz hám onıń qollanıwı oǵada áhmiyetli bolıp tabıladi. Atap

ótsek, ilimiy izertlewler ushın qollanǵanda korpus ádebiyattanıwshılar, lingvistler hám basqa da izertlewshiler ushın bahalı resurs bolıp esaplanadı. Ol T.Qayıpbergenovtıń shıǵarmaların hár tárepleme analizlewge, onıń stilin, til ózgesheliklerin hám tematikalıq baǵdarların tereń úyreniwge múmkinshilik beredi. Al, bilimlendiriw maqsetleri ushın paydalanǵanda bolsa, korpus mektep hám joqarı oqıw orınlarında qaraqalpaq tili hám ádebiyatın oqıtıwda qollanıwı múmkin. Ol studentlerge T.Qayıpbergenovtıń shıǵarmaları menen jaqınnan tanıwǵa hám olardı hár tárepleme úyreniwge járdem beredi. Til úyreniw ushın bolsa, álbette, korpus qaraqalpaq tilin úyreniwshilerge paydalı resurs bolıp tabıladi. Ol tildiń haqıyqıy qollanıwın kórsetedi hám sózlik qorın bayıtıwǵa járdem beredi. Usı arqalı mádeniy miyrastı saqlap, korpus T.Qayıpbergenovtıń shıǵarmaların sanlı formatta ámelge asırıwǵa hám olardı keleshek áwladlarǵa jetkeriwge múmkinshilik beredi. Eń baslısı, qaraqalpaq tilin jáne de rawajlandırıw. Korpus qaraqalpaq tiliniń kompyuter lingvistikası tarawında qollanıwın keńeytedi, bul tildiń rawajlanıwına hám saqlanıwına úlken úles qosadı.

Juwmalaqap aytqanda, T.Qayıpbergenovtıń shıǵarmalar korpusı, bul – qaraqalpaq tili hám ádebiyatı boyınsha izertlewler ushın áhmiyetli resurs. Ol jazıwshınıń miyrasın saqlawǵa, úyreniwge hám keń jámiyetshilikke jetkeriwge úlken úles qosadı. Korpusıń rawajlanıwı hám keńeyiwi qaraqalpaq tilin hám ádebiyatın izertlewdiń jańa múmkinshiliklerin ashadı, sonday-aq, ulıwma qaraqalpaq mádeniyatı hám tariyxın úyreniwge de járdem beredi. Korpus tek ǵana ilimpazlar ushın emes, al ulıwma qaraqalpaq tili hám ádebiyatına qızıǵıwshı hár bir adam ushın da paydalı bolıwı múmkin. Ol T.Qayıpbergenovtıń shıǵarmaların jańasha túsiniwge, onıń stilin hám til baylıǵın tereń ańlawǵa múmkinshilik beredi. Keleshekte bul korpus qaraqalpaq tiliniń rawajlanıwına, onıń sanlı texnologiyalar dáwirinde saqlanıwına hám keń qollanıwına tiykar bolıwı múmkin. Sonlıqtan, onı rawajlandırıw hám jetilistiriw – bul tek ǵana ilimiy maqset emes, al ulıwma qaraqalpaq mádeniyatın saqlaw hám rawajlandırıw boyınsha áhmiyetli wazıypa bolıp esaplanadı.

### Ádebiyatlar

1. McEnery T., & Hardie A. 2011. Corpus linguistics: Method, theory and practice. Cambridge University Press.
2. Sinclair J. 2004. Trust the text: Language, corpus and discourse. Routledge.
3. Dash N.S. 2008. Corpus Linguistics: An Introduction. Pearson Education India.

**РЕЗЮМЕ.** Маққала қарақалпақ тілі асосында лингвистик корпус қаратыш жарыонини қорытышқа қаратылған. Unda T.Qayıpbergenov асарлары мисолыда тілні тизимлі тahlil қилыш усулары ва ахборот тизимлары имкониятлары ко'рсатып о'тилған. Табиий тілні қайта ишлалаш технологиялары ва лингвистик корпусның аһмийаты тилшунослик соһасыда инновациян қондашув сифатыда тahlil қилыған. Маққала корпус қаратыш босқичларини, техник јihatларини ва уның илмий-амалий аһмийатыни байон етadi.

**РЕЗЮМЕ.** В статье рассматривается процесс создания лингвистического корпуса на основе каракалпакского языка. На примере произведений Т.Қайыпбергенова представлены методы системного анализа языка и возможности информационных систем. Технология обработки естественного языка и значение лингвистического корпуса анализируются как инновационный подход в области лингвистики. Также описаны этапы создания корпуса, его технические аспекты и научно-практическое значение.

**SUMMARY.** This article focuses on the process of creating a linguistic corpus based on the Karakalpak language. Using T.Kaipbergenov's works as an example, it highlights methods for systematic language analysis and the potential of information systems. The importance of natural language processing technologies and linguistic corpora is analyzed as an innovative approach in linguistics. The article outlines the stages of corpus creation, its technical aspects, and its scientific and practical significance.