QORAQALPOQ IDIOMALARINI RAQAMLI MUHITDA NAShR ETIShNING LINGVISTIK VA METODOLOGIK ASOSLARI

Saparov Salamat Perdebaevich – *filologiya fanlari bo‘yicha falsafa doktori*
ssaparov202581@gmail.com
Masuatdinova Inkar Sultanovna – *magistrant*
Ajiniyoz nomidagi Nukus davlat pedagogika instituti

ЛИНГВИСТИЧЕСКИЕ И МЕТОДОЛОГИЧЕСКИЕ ОСНОВЫ ПРЕДСТАВЛЕНИЯ КАРАКАЛПАКСКИХ ИДИОМ В ЦИФРОВОЙ СРЕДЕ

Сапаров Саламат Пердебаевич – *доктор философии по филологическим наукам*
Масуатдинова Инкар Султановна – *магистрант*
Нукусский государственный педагогический институт имени Ажинияза

LINGUISTIC AND METHODOLOGICAL FOUNDATION FOR PRESCULITY KARAKALPAK IDIOMS IN ADIGITAL ENVIROUMENT

Saparov Salamat Perdebaevich – *Doctor of Philosophy in Philological sciences*
Masuatdinova Inkar Sultanovna – *master's student*
Nukus State Pedagogical Institute named after Ajiniyaz

Tayanch so‘zlar: qoraqalpoq idiomalari, raqamli leksikografiya, frazeologiya, lingvistik korpus, onlayn lug‘at, tilni saqlash, veb texnologiyalar, kam sonli tillar.

Rezyume. Maqolada qoraqalpoq idiomalarini raqamli platformalarda nashr etishning asosiy yondashuvlari tahlil qilinadi. Tadqiqotda raqamli leksikografiya, korpus lingvistikasi, multimedia texnologiyalari hamda veb-resurslardan foydalanish imkoniyatlari ko‘rib chiqilib, ularning qoraqalpoq tilining frazeologik boyligini saqlash, ommalashtirish va keng jamoatchilikka yetkazishdagi o‘rni yoritiladi.

Ключевые слова: каракалпакские идиомы, цифровая лексикография, фразеология, лингвистический корпус, онлайн-словарь, сохранение языка, веб-технологии, языки малочисленных народов.

Резюме. В статье рассматриваются основные подходы к публикации каракалпакских идиом на цифровых платформах. Анализируются возможности цифровой лексикографии, корпусной лингвистики, мультимедийных технологий и веб-ресурсов, направленных на сохранение, систематизацию и распространение фразеологического наследия каракалпакского языка.

Key words: Karakalpak idioms, digital lexicography, phraseology, linguistic corpus, online dictionary, language preservation, web technologies, minority languages.

Summary. This article discusses the principal approaches to publishing Karakalpak idioms on digital platforms. It analyzes digital lexicography, corpus-based methods, multimedia technologies, and web resources aimed at preserving and disseminating Karakalpak linguistic heritage.

Introduction. Karakalpak is a Turkic language of the Kipchak (Kipchak–Nogai) group, spoken primarily in the Republic of Karakalpakstan, a sovereign republic within the Republic of Uzbekistan, and by smaller Karakalpak communities in Kazakhstan, Turkmenistan, Russia and other parts of Central Asia. Estimates of the number of speakers vary across sources and depend on the criteria used, so contemporary figures should be treated with caution. Despite its established sociolinguistic status and its use in education, the media and public communication, Karakalpak remains under-resourced in language technology: openly accessible corpora, electronic dictionaries and natural language processing tools for the language are still limited. This situation calls for the systematic development of digital resources for the preservation, documentation and further development of the language in the digital environment.

The Karakalpak language possesses a rich stock of idioms and phraseological units that reflect the worldview of its speakers, their history in the Amu Darya delta and their close connection with nature, livestock breeding and fishing. The grammatical and lexical system of the language was described in the foundational works of N.A.Baskakov [1] and E. Berdimuratov [3]; Karakalpak phraseology has been studied by D.S.Nasyrov [7] and A.Bekbergenov [2], and its units have been collected in a phraseological dictionary [6]. The general theory of phraseology was developed in the Russian tradition by V.V.Vinogradov [4] and V.N.Teliya [8] and, for English, by A.V.Kunin [5], A.P.Cowie [9] and R. Moon [13]. Since idioms encode the cultural and historical experience of a people [8], and the loss of a language entails the loss of the cultural and ecological knowledge it carries [10], the phraseological heritage of Karakalpak requires documentation that meets the standards of documentary linguistics [11] and is accessible to a global audience.

Publishing Karakalpak idioms online is therefore not merely a matter of digitizing a printed dictionary. It means creating an open, updatable resource that serves researchers, teachers and learners and that links Karakalpak phraseology with other languages. For students of English philology in Karakalpakstan, systematically presented Karakalpak–English idiom correspondences are also a practical resource for translation and language teaching.

The *aim* of this study is to substantiate a set of approaches to the digital publication of Karakalpak idioms. To achieve this aim, the following *objectives* were set: (1) to review the lexicographic and theoretical literature on Karakalpak and general phraseology; (2) to compare a sample of Karakalpak idioms with their English equivalents and identify the difficulties of rendering them; (3) to analyse existing models of digital lexicography and

endangered-language documentation; and (4) to propose a model for publishing Karakalpak idioms on the World Wide Web.

The *scientific novelty* of the study lies in proposing, for the first time for Karakalpak, a comprehensive model for the digital publication of idioms that integrates lexicographic, corpus-based, multimedia and web technologies with a Karakalpak–English contrastive component.

Methodology. The study follows a qualitative, descriptive-analytical design combining four complementary methods. First, a review of Karakalpak linguistic and phraseological scholarship was carried out, drawing on the works of N.A.Baskakov [1], E.Berdimuratov [3], K.M.Maksetov [6] and D.S.Nasyrov [7], the semantic classification proposed by A.Bekbergenov [2], and the general theory of phraseology developed by V.V.Vinogradov [4], A.V.Kunin [5], V.N.Teliya [8], A.P.Cowie [9] and R.Moon [13]. Second, a comparative-contrastive analysis was performed on a purposive sample of ten Karakalpak idioms, each matched with its closest English functional equivalent (Table 1), in order to identify translation strategies and to test the adequacy of literal versus functional rendering. Third, a benchmarking analysis of existing models of digital lexicography and language documentation – the Endangered Languages Archive (ELAR), the Endangered Archives Programme of the British Library, the FLEx (FieldWorks Language Explorer) lexical database format [11; 14] and the OntoLex-Lemon model for linked lexical data [12] – was undertaken to identify technical solutions transferable to the Karakalpak context. Fourth, the findings of these three strands were synthesized through thematic classification, grouping recurring lexicographic, technical and sociocultural strategies into the nine approaches reported below.

The design moves from linguistic description – what the idioms mean and why they resist word-for-word translation – to applied recommendation – how they can be documented, structured and disseminated online. It is intended as a conceptual and comparative framework rather than an empirical evaluation of a deployed system; this limitation is discussed below.

Results. The comparative analysis of the sample (Table 1) shows that most Karakalpak idioms cannot be adequately rendered into English word for word: only in isolated cases (e.g., No. 3) do the two languages share the same image, while in the majority of cases a functional equivalent based on a different image is required.

Table 1. Sample of Karakalpak idioms and their English equivalents

Karakalpak idiom	English equivalent	Meaning
qulaǵın tartıw	to pull someone's leg	to tease, to fool someone
tildi tistew	to hold one's tongue	to keep silent
it sıyaqlı islew	to work like a dog	to work very hard
jandı tapsırıw	to kick the bucket	to die
pullı bolıw	to be rolling in money	to be very rich
jaqa jırtısıw	to be at each other's throats	to quarrel fiercely
kókte ushıw	to be on cloud nine	to be overjoyed
bası qatıw	to rack one's brains	to think hard; to be puzzled
túye murnınan jip ótkiziw	to twist someone around one's finger	to control or dominate someone
júregi awzına tıǵılıw	to have one's heart in one's mouth	to be very frightened or anxious

On the basis of this analysis and of the literature and model review, nine principal approaches to the digital publication of Karakalpak idioms were identified.

1. *From dictionaries to linguistic databases.* Printed dictionaries and descriptions [1; 3; 6] remain the foundation, but idioms must function differently online. The semantic groups of Karakalpak phraseological units already identified in research [2] provide a basis for structuring a digital database. Each entry should store not only the headword and its definition but also a literal gloss, a functional equivalent in other languages and authentic usage examples. This is essential because many Karakalpak idioms cannot be translated word for word: a literal rendering of *bası qatıw* (lit. ‘one’s head hardens’) would be meaningless to an English speaker, whereas the functional equivalent *to rack one’s brains* conveys the intended sense.

2. *Digital lexicography and online dictionaries.* One of the most effective approaches is the development of an online dictionary of Karakalpak idioms [14]. Such a resource, once openly published, can also serve as reliable reference data for machine translation and other language technologies. Each entry should include:

– a definition in Karakalpak and translations into widely spoken languages (English, Russian, Uzbek);

– contextual examples from literature and the media;

– phonetic transcription and stress marking.

Digital lexicography allows continuous updating of entries and flexible search, making idioms easily accessible to users [12; 14].

3. *The corpus approach.* A national corpus of the Karakalpak language would allow users to see idioms in real sentences from literature and the media, helping learners grasp nuances of usage that are lost in literal translation. The corpus-based approach to fixed expressions and idioms elaborated by R.Moon [13] offers a reliable methodological model for Karakalpak.

4. *Multimedia integration.* Publishing idioms with multimedia elements enhances understanding and engagement. This includes:

– audio recordings by native speakers;

– video explanations and storytelling;

– visual illustrations of idiomatic meanings;

– short animated videos covering thematic groups of idioms.

Multimedia resources are particularly useful for language learners and help preserve pronunciation and oral tradition.

5. *Technical implementation.* Online lexicographic databases are typically built on relational database

management systems (MySQL, PostgreSQL) with a web front-end (HTML, CSS, JavaScript) and a server-side language such as PHP or Python. A RESTful API layer makes the data accessible programmatically to researchers, application developers and NLP specialists. To make entries interoperable and discoverable, they can be modelled according to the W3C OntoLex-Lemon vocabulary [12] and annotated with Schema.org structured data (e.g., the *DefinedTerm* type) in JSON-LD. All Karakalpak characters (á, ó, ú, ı, ǵ, ń) must be stored in UTF-8 Unicode. Established models include the Endangered Languages Archive (ELAR), the online collections of the British Library's Endangered Archives Programme and the FLEx (FieldWorks Language Explorer) format used by many field linguists [11; 14].

6. *Challenges and solutions.* The main challenges are:

– limited digital resources;

– lack of funding and technical expertise;

– risk of misinterpretation when idioms are presented without context.

Solutions include partnerships with universities, open-source platforms, and governmental and cultural initiatives supporting minority languages. The existing documentation of Karakalpak phraseology [6; 7] already provides textual material for building an initial database. Creating a high-quality lexicographic resource also requires sustained editorial oversight, trained lexicographers and native-speaker consultants.

7. *Web platforms and mobile applications.* Dedicated websites and mobile applications provide interactive ways to explore idioms. Key features may include:

– daily idiom notifications;

– quizzes and gamified learning;

– search and categorization by theme (nature, emotions, work, etc.).

Mobile accessibility ensures wider reach, especially among young people, which is essential for maintaining the vitality of the language.

8. *Social media and community engagement.* Social media play a crucial role in popularizing idioms. Possible strategies include posting "Idiom of the Day" content, encouraging user-generated examples and creating discussion groups for language learners. This participatory approach fosters community involvement and keeps the language in active use.

9. *Standardization and linguistic accuracy.* To ensure quality, it is essential to involve native speakers as well as specialists in Karakalpak and English linguistics [14], to standardize spelling and meanings, and to indicate the source of every entry. Three further components strengthen this approach:

– *comparative databases* – parallel lists in which Karakalpak idioms are matched with English, Uzbek and Russian equivalents, following the principles of contrastive phraseology [5; 9];

– *etymological notes* – short comments on the origin of an expression (for example, its connection with nomadic life, livestock breeding or fishing), which make the content engaging for a global audience and preserve the cultural knowledge embedded in it [8];

– *alphabet and orthography* – Karakalpak has been written in Arabic-based, Latin-based (from 1928) and Cyrillic (from 1940) scripts, and since 1994 in a new Latin alphabet that was subsequently revised. A digital resource should therefore store each idiom in the current official Latin orthography together with its Cyrillic form, and its search function should tolerate spelling variants (e.g., *g‘* / *ǵ*, *n‘* / *ń*).

Discussion. The comparative sample in Table 1 clarifies why word-for-word rendering, still typical of printed bilingual glossaries, is inadequate for a digital audience unfamiliar with Karakalpak imagery. An idiom built on a local referent, such as *túye murnınan jip ótkiziw* (lit. 'to pass a rope through a camel's nose'), requires the contextual and multimedia support described in Approaches 3 and 4 rather than a single-line gloss. This confirms, for Karakalpak specifically, the view in the phraseological literature [8; 9; 13] that idiomaticity operates at the level of discourse and cultural knowledge rather than at the level of the individual word.

The difficulties identified under Approach 6 – limited funding, scarce technical expertise and the risk of decontextualized interpretation – mirror obstacles widely reported for other minority-language documentation projects [10, 11, 14]. This suggests that the barriers facing Karakalpak digitization are structural rather than language-specific, and that solutions tested in comparable projects, such as university partnerships and open-source tools, can be transferred to the Karakalpak context rather than invented anew.

At the same time, the scope of the present study is limited. The proposed model rests on literature-based and comparative analysis rather than on the development and user testing of an actual digital platform; the idiom sample in Table 1 (n = 10) was selected for illustration rather than through systematic corpus sampling; and the orthographic issues raised under Approach 9 have not been tested against real user search behaviour. Future research should therefore complement this conceptual model with a pilot online dictionary and usability testing among learners and native speakers, so that the proposed approaches can be validated empirically.

Conclusion. Publishing Karakalpak idioms on the World Wide Web is an important step towards preserving the linguistic heritage of the Karakalpak people and making it accessible to a global audience [10; 14]. The study has shown that Karakalpak idioms require functional rather than literal rendering in other languages and has proposed nine interrelated approaches to their digital publication: a structured lexicographic database, an online dictionary, a corpus component, multimedia support, an interoperable technical architecture, institutional solutions to resource constraints, web and mobile platforms, community engagement, and standardization of spelling and content. The practical value of the model lies in its applicability to the creation of an online Karakalpak–English (–Russian–Uzbek) dictionary of idioms and to the teaching of English and Karakalpak phraseology at university level. Further work should focus on building a pilot platform, expanding the idiom database through corpus sampling and evaluating the resource with its users.

References

1. Баскаков Н.А. Каракалпакский язык. Т. II: Фонетика и морфология. Ч. 1. – Москва: Изд-во АН СССР, 1952.
2. Бекбергенов А. Қарақалпақ тилиндеги фразеологиялық бирликлердиң семантикалық топарлары. // Өзбекстан Республикасы Илимлер академиясы Қарақалпақстан бөлиминиң Хабаршысы, 2003, № 2. 45-52-б.
3. Бердимуратов Е. Ҳәзирги қарақалпақ тили. Лексикология. – Нөкис: «Билим», 1994.
4. Виноградов В.В. Об основных типах фразеологических единиц в русском языке. Избранные труды: Лексикология и лексикография. – Москва: «Наука», 1977. – С. 140-161.
5. Кунин А.В. Курс фразеологии современного английского языка. 2-е изд., перераб. – Москва: «Высшая школа»; – Дубна: «Феникс», 1996.
6. Мақсетов Қ.М. Қарақалпақ тилиниң фразеологиялық сөзлиги. – Нөкис: «Қарақалпақстан», 1994.
7. Насыров Д.С. Қарақалпақ тилиниң фразеологиясы. – Нөкис: «Қарақалпақстан», 1982.
8. Телия В.Н. Русская фразеология: Семантический, прагматический и лингвокультурологический аспекты. – Москва: Школа «Языки русской культуры», 1996.
9. Cowie A.P. (ed.). Phraseology: Theory, Analysis, and Applications. – Oxford: Clarendon Press, 1998.
10. Harrison K.D. When Languages Die: The Extinction of the World's Languages and the Erosion of Human Knowledge. – Oxford: Oxford University Press, 2007.
11. Himmelmann N.P. Documentary and descriptive linguistics. // Linguistics, 1998, Vol. 36, № 1. – P. 161-195.
12. McCrae J. P., Bosque-Gil J., Gracia J., Buitelaar P., Cimiano P. The OntoLex-Lemon Model: Development and Applications. // Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference. – Brno: Lexical Computing CZ, 2017. – P. 587-597.
13. Moon R. Fixed Expressions and Idioms in English: A Corpus-Based Approach. – Oxford: Clarendon Press, 1998.
14. Ogilvie S. Linguistics, Lexicography, and the Revitalization of Endangered Languages. // International Journal of Lexicography, 2011, Vol. 24, № 4. – P. 389-404.