**MRNTI: 16.41.25**
**DOI: https://doi.org/10.66941/3106-9037/2026/iss3pp68-76**

**Mirzayev F.R.*, Jafarova G.K.**

*Ganja Branch of the Azerbaijan National Academy of Sciences, 419 Heydar Aliyev Avenue, AZ 2003, Ganja, Azerbaijan*

## ARTIFICIAL INTELLIGENCE FOR THE DIGITAL PRESERVATION OF GANJA-GAZAKH-KARABAKH DIALECTS

**Mirzayev F.R.,** Master's degree, Head of the Department of Artificial Intelligence and Digitalization;
E-mail: grem2011@gmail.com; ORCID id: 0009-0006-6303-2343
**Jafarova G.Kh.,** Master's degree, Scientific Secretary;
E-mail: gunayceferova.0125@gmail.com; https://orcid.org/ : 0000-0001-6990-4305

**Abstract** .The article examines the digital transformation of the Ganja-Gazakh-Karabakh dialects and patois of the Azerbaijani language in the context of artificial intelligence (AI) technologies. The purpose of the study is to develop a conceptual model that reconciles the preservation of dialectal diversity with the development of the national digital language infrastructure. The tasks include determining the structural requirements of a dialect corpus, designing an institutional cooperation model, systematizing AI applications, and analyzing the ethical contradictions between standardization and diversity. Comparative-historical, systemic and content-analysis methods and theoretical generalization were applied to scholarly literature and international documents. The results show that annotated corpora enriched with sociolinguistic metadata are a precondition for dialect-oriented AI technologies, while systematic cooperation among the state, academic institutions and the technology sector ensures their sustainability. It is concluded that inclusive AI-based language technologies can shift dialects from archiving to active everyday use.

**Keywords:** Azerbaijani language, Ganja-Gazakh-Karabakh dialects, dialect preservation, artificial intelligence, digital transformation, language policy, corpus linguistics.

**Acknowledgments.** The work was carried out within the institutional research plan of the Ganja Branch of the Azerbaijan National Academy of Sciences.

For citation: Mirzayev F.R., Jafarova G.Kh. Artificial intelligence for the digital preservation of Ganja-Gazakh-Karabakh dialects / Journal "Science of Intelligent Systems". 2026. Vol. [2]. No. [2]. P. [9-20]. Language of the article: English; (issue [2]), https://doi.orghttps:// 10.66941/3106-9037/2026/iss2pp9-20

**Ф.Р. Мирзаев*, Г.Х. Джафарова**

*Гянджинское отделение Национальной академии наук Азербайджана, проспект Гейдара Алиева, 419, AZ 2003, г. Гянджа, Азербайджан*

## ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ЦИФРОВОМ СОХРАНЕНИИ ГЯНДЖА-ГАЗАХСКО-КАРАБАХСКИХ ДИАЛЕКТОВ

**Мирзаев Ф.Р.,** магистр, заведующий отделом искусственного интеллекта и цифровизации;
E-mail: grem2011@gmail.com; https://doi.org/ 0009-0006-6303-2343
**Джафарова Г.Х.,** магистр, научный секретарь;
E-mail: gunayceferova.0125@gmail.com; https://doi.org/ 0000-0001-6990-4305

**Аннотация**. В статье рассматривается цифровая трансформация Гянджа-Газахско-Карабахских диалектов и говоров азербайджанского языка в контексте технологий искусственного интеллекта (ИИ). Цель исследования – разработать концептуальную модель, согласующую сохранение диалектного многообразия с развитием национальной цифровой языковой инфраструктуры. Задачи включали определение структурных требований к диалектному корпусу, проектирование модели институционального сотрудничества, систематизацию направлений применения ИИ и анализ этических противоречий между стандартизацией и многообразием. Использованы сравнительно-исторический метод, системный анализ, контент-анализ и теоретическое обобщение научной литературы и международных документов. Показано, что аннотированные корпуса с социолингвистическими метаданными служат необходимым условием для ИИ-технологий, ориентированных на диалекты, а системное сотрудничество государства, науки и технологического сектора обеспечивает устойчивость процесса. Сделан вывод, что инклюзивные языковые технологии на основе ИИ позволяют перейти от архивирования диалектов к их активному повседневному использованию.

**Ключевые слова:** азербайджанский язык, гянджа-газахско-карабахские диалекты, сохранение диалектов, искусственный интеллект, цифровая трансформация, языковая политика, корпусная лингвистика.

**Благодарность.** Работа выполнена в рамках институционального плана исследований Гянджинского отделения Национальной академии наук Азербайджана.

**Для цитирования:** Мирзаев Ф.Р., Джафарова Г.Х. Искусственный интеллект в цифровом сохранении гянджа-газахско-карабахских диалектов / Журнал «Наука интеллектуальных систем». 2026. Т. [2_]. № [2]. С. [9-20]. Язык статьи: английский; (номер журнала [2]), https://doi.org/ 10.66941/3106-9037/2026/iss2pp9-20

**Ф.Р. Мирзаев*, Г.Х. Джафарова**

*Әзербайжан Ұлттық ғылым академиясының Гянджа бөлімшесі, Гейдар Әлиев даңғылы, 419, AZ 2003, Гянджа қ., Әзербайжан*

## ГЯНДЖА-ГАЗАХ-ҚАРАБАҚ ДИАЛЕКТІЛЕРІН ЦИФРЛЫҚ САҚТАУДАҒЫ ЖАСАНДЫ ИНТЕЛЛЕКТ

**Мирзаев Ф.Р.,** магистр, жасанды интеллект және цифрландыру бөлімінің меңгерушісі;
E-mail: grem2011@gmail.com; https://doi.org/ 0009-0006-6303-2343
**Джафарова Г.Х.,** магистр, ғылыми хатшы;
E-mail: gunayceferova.0125@gmail.com; https://doi.org/ 0000-0001-6990-4305

**Аннотация.** Мақалада жасанды интеллект (ЖИ) технологиялары контекстінде әзербайжан тілінің Гянджа-Газах-Қарабақ диалектілері мен говорларын цифрлық трансформациялау мәселелері қарастырылады. Зерттеудің мақсаты – диалектілік алуантүрлілікті сақтауды ұлттық цифрлық тілдік инфрақұрылымды дамытумен үйлестіретін тұжырымдамалық модель әзірлеу. Міндеттерге диалект корпусына қойылатын құрылымдық талаптарды айқындау, институционалдық ынтымақтастық моделін жобалау, ЖИ қолдану бағыттарын жүйелеу және стандарттау мен алуантүрлілік арасындағы этикалық қайшылықтарды талдау кірді. Ғылыми әдебиет пен халықаралық құжаттарға салыстырмалы-тарихи әдіс, жүйелік талдау, контент-талдау және теориялық жалпылау қолданылды. Әлеуметтік-лингвистикалық метадеректермен аннотацияланған корпустар диалектілерге бағытталған ЖИ технологиялары үшін қажетті алғышарт екені, ал мемлекет, ғылым және технология секторы арасындағы жүйелі ынтымақтастық үдерістің тұрақтылығын қамтамасыз ететіні көрсетілді. Инклюзивті ЖИ негізіндегі тілдік технологиялар диалектілерді мұрағаттаудан оларды белсенді күнделікті қолдануға көшуге мүмкіндік береді деген қорытынды жасалды.

**Түйін сөздер:** әзербайжан тілі, Гянджа–Газах–Қарабақ диалектілері, диалектілерді сақтау, жасанды интеллект, цифрлық трансформация, тіл саясаты, корпустық лингвистика.

**Алғыс.** Жұмыс Әзербайжан Ұлттық ғылым академиясы Гянджа бөлімшесінің институционалдық зерттеу жоспары шеңберінде орындалды.

**Дәйексөз үшін:** Мирзаев Ф.Р., Джафарова Г.Х. Гянджа-Газах-Қарабақ диалектілерін цифрлық сақтаудағы жасанды интеллект / «Наука интеллектуальных систем» журналы. 2026. Т. [2]. № [2]. Б. [9-20]. Мақала тілі: ағылшын; (журнал нөмірі [2]), https://doi.org/ 10.66941/3106-9037/2026/iss2pp9-20

**Introduction**

The rapid development of digital technologies has fundamentally changed approaches to the preservation, documentation, and use of languages. For the Azerbaijani language, rich in Ganja-Gazakh-Karabakh dialects and patois, digital transformation creates both new opportunities and serious challenges. While artificial intelligence (AI) technologies make it possible to process, classify, and generate language material on a large scale, they simultaneously strengthen the dominance of standard language forms, increasing the risk that regional variants will be pushed into the shadows. This situation places a multifaceted task before researchers: to benefit from technological progress while at the same time preserving linguistic diversity.

The dialectological map of the Azerbaijani language is complex and rich; the Ganja-Gazakh-Karabakh group is one of its historically, phonetically, and lexically most distinctive branches (Shiraliyev, 2008; Akhundov, 1984; Hamzayev, 1992). These dialects are not merely an object of linguistic study but an inseparable part of the region's cultural life, traditions, and collective memory. Folklore examples – tales, bayati verses, laments, riddles – are among the principal means of carrying and sustaining the vitality of these dialects. The digital era is not content merely with archiving such materials; it creates additional opportunities to turn them into a living, actively used language resource.

The relevance of the topic is conditioned by several factors. First, the rate of loss of languages and dialects worldwide is increasing, and according to UNESCO, thousands of languages and dialects are endangered (UNESCO, 2010). Second, although the rapid development of AI technologies is expanding the possibilities for processing language resources, the question of who will benefit from these opportunities remains open: if dialect materials are not represented in the digital environment, they will remain outside the "field of vision" of future AI models. Third, the socio-economic processes taking place in the Ganja-Gazakh-Karabakh region, including urbanization and the weakening of intergenerational language transmission, are leading to a decrease in the number of living carriers of the dialects.

This article considers the digital transformation of the Ganja-Gazakh-Karabakh dialects as a strategic issue of linguistics and cultural heritage. The aim of the study is to develop a conceptual model that reconciles the preservation of dialectal diversity with the development of the national digital language infrastructure. The object of the study is the totality of technological, institutional, and ethical factors conditioning the digital transformation of the Ganja-Gazakh-Karabakh dialects and patois; its subject is the process of this transformation. To achieve the aim, the following tasks were set: determining the structural requirements of the dialect corpus; designing an institutional cooperation model; systematizing the possibilities for applying AI technologies in the dialect environment; and conducting an ethical analysis of the contradictions between standardization and diversity. The hypothesis of the study is that the digitalization of dialect resources within an institutional framework, in accordance with ethical principles, can create a sustainable model ensuring both their preservation and their active use. The study relies on the approaches of corpus linguistics, computational sociolinguistics, and documentary linguistics, and uses the comparative-historical method, systemic analysis, content analysis, and theoretical generalization.

The scientific novelty of the study lies in the fact that, for the first time, the process of digital transformation of the Ganja-Gazakh-Karabakh dialect group is analyzed within a unified conceptual framework – a synthesis of the resource, institutional cooperation, and ethical dimensions. The theoretical significance of the work lies in the fact that it synthesizes knowledge accumulated in the fields of corpus linguistics, computational sociolinguistics, and documentary linguistics on the basis of the material of a specific regional dialect group, thereby proposing a model for the practical application of theoretical generalizations. In terms of practical significance, the results of the work can serve to

define concrete lines of action for state institutions, scientific establishments, and technology companies.

The structure of the article is as follows. The section "Research materials and methods" describes the methodology and the material of the study. The section "Discussion" reviews the scholarly literature and international experience. The section "Results" analyzes, in sequence, the structural requirements of the dialect corpus, the institutional cooperation model, AI applications, standardization and ethical challenges, lexical-semantic aspects, risks, and the strategy of integration into education, and then summarizes the findings. The "Conclusion" sums up the work and identifies directions for future research.

**Research materials and methods Methodology**

The methodological basis of the study is formed by the scientific approaches developed in the fields of corpus linguistics (McEnery & Hardie, 2012), computational sociolinguistics (Nguyen et al., 2016), and documentary linguistics (Himmelmann, 1998). The following methods were used: the comparative-historical method (comparison of the Ganja-Gazakh-Karabakh dialects with dialects of other Turkic languages); the systemic analysis method (modeling institutional links among the state, science, and technology sectors); the content-analysis method (analysis of the existing scholarly literature and international documents); and the method of theoretical generalization.

The research questions are as follows. What structural requirements must a dialect corpus meet in order to serve as a basis for AI applications? What institutional model can ensure the sustainability of digital dialect projects? Which AI technologies can move dialects from archiving to active use? How can the contradictions between standardization and dialectal diversity be resolved ethically? The hypothesis of the study is formulated in the Introduction.

The study was carried out in five stages: (1) selection of sources and identification of the problem field; (2) content analysis of the scholarly literature and international documents; (3) comparative analysis of international experience with low-resource languages and dialects; (4) modeling of the institutional cooperation scheme and systematization of AI applications; (5) ethical analysis and theoretical generalization of the findings.

**Material of the study**

The material of the study is qualitative in nature. It comprises twenty sources listed in the References: six works in Azerbaijani (including the dialectological studies of Shiraliyev, Akhundov, and Hamzayev, and the materials on the national corpus published by the Institute of Linguistics of the Azerbaijan National Academy of Sciences (ANAS)) and fourteen international works, including UNESCO's atlas of endangered languages and descriptions of similar projects in world practice, in particular AI tools developed for low-resource languages and dialects.

The study is based on published sources; no new primary field data were collected at this stage. The novelty in the use of the material consists in bringing together dialectological, computational, and sociolinguistic sources, which have so far been considered separately, in the analysis of one regional dialect group.

**Discussion**

The topic of the digital preservation of dialects has gained particular importance in international linguistics in recent decades. Jurafsky and Martin (2023), systematizing the methods of speech and language processing, describe approaches whose quality depends heavily on the volume of training data; this dependence creates specific difficulties for low-resource language variants such as dialects. Nguyen et al. (2016), in their survey of computational sociolinguistics, show that language use can be studied in relation to social variables (age, gender, region) by computational methods, which substantiates the importance of enriching dialect materials with metadata.

The concept of the "Bender Rule", put forward by Bender (2019), emphasizes the necessity of explicitly stating the name and characteristics of the language in the development of language technologies; this approach is important for preventing dialect diversity from remaining "invisible" in AI models. Himmelmann (1998) substantiates documentary linguistics as a field aimed at creating a lasting, multipurpose record of a language, which forms one of the methodological foundations of the present study. McEnery and Hardie (2012), describing the methodological tools of corpus linguistics, show the direct impact of balanced and annotated corpora on the quality of language technologies.

In the local scholarly literature, Shiraliyev's (2008) work "Fundamentals of Azerbaijani Dialectology" is a fundamental source providing a systematic description of the phonetic and morphological features of the Ganja-Gazakh-Karabakh dialect group. Akhundov (1984) and Hamzayev (1992) further deepened this direction, proposing additional criteria for the classification of dialects and patois. Alizadeh's (2016) work "Dialects of Western Azerbaijan" is significant from the point of view of comparative analysis of neighboring dialect groups. UNESCO's (2010) atlas, by documenting the pace of loss of dialects and less-spoken languages worldwide, substantiates the necessity of taking urgent measures. In this context, the work carried out by the Institute of Linguistics of ANAS (2021) on the national corpus project is an indicator of the institutional foundation forming within the country in this direction.

In world practice, there are a number of successful projects in the field of digital preservation of low-resource languages and dialects. Most of them are based on three common principles: community-based data collection, the use of open-source software, and compliance with long-term archiving standards (Hinton, Huss & Roche, 2018). The community-based approach means that dialect speakers are engaged as active participants in the project rather than as passive sources of information; this both increases the quality of the data and strengthens the community's sense of ownership over its own language.

Grenoble and Whaley (2006), in their systematic analysis of language revitalization, show that the main feature of successful revitalization projects is that technology is applied not as a solution on its own but as part of a broader social and educational strategy. Crystal (2000), analyzing the social mechanisms of language death, emphasizes that the interruption of intergenerational transmission is the main risk factor; from this point of view, digital tools that attract the attention of the younger generation (mobile applications, social media integration) are of particular importance. Trudgill (2000), examining the mechanisms of sociolinguistic change, shows that the relationship between the standard language and dialects is closely tied to social prestige. Labov (2001), in turn, substantiates through empirical research that language change is conditioned by social factors, which indicates the importance of taking social variables (age, education, place of residence) into account in the design of digital tools.

Bauman and Briggs (1990), within the framework of the theory of verbal performance, note the importance of documenting not only the linguistic but also the performative characteristics of folklore texts – genres such as tales, bayati verses, and laments. Bucholtz and Hall (2005), within the framework of the theory of identity and interaction, analyze the role played by language choice in the construction of social identity; this approach shows the importance of evaluating dialect use not merely as a linguistic fact but as an expression of social identity. Thus, the general lesson drawn from international experience is that technological solutions produce more sustainable results when linked to sociolinguistic theory.

Nevertheless, the number of systematic works in the existing literature on the integration of the rich dialectological heritage of the Ganja-Gazakh-Karabakh region with AI technologies remains limited. The Azerbaijani dialectological studies describe the dialects but do not address their digital representation, while the international works on low-resource languages do not take into account the specific features of this dialect group. The present study fills this gap.

**Results Structure and technological foundations of the dialect corpus**

One of the decisive factors in the digital transformation of the Azerbaijani language is the creation of large-scale, structured, and high-quality digital language resources (McEnery & Hardie, 2012). Text corpora, Ganja-Gazakh-Karabakh dialect materials, and audio archives play the role of the basic infrastructure for AI applications such as automatic speech recognition, dialect identification, and text generation (Jurafsky & Martin, 2023).

The dialect corpus should consist not only of old written texts but also of modern spoken language, folklore examples (tales, laments, riddles, bayati verses), as well as sociolinguistic interviews. The phonetic, morphological, and lexical features of the dialect should be annotated in detail within the corpus. Metadata related to each text or audio recording (the age, gender, place of residence, occupation, and language environment of the information source) should be collected; this provides valuable data for studying the social dynamics of language change (Nguyen et al., 2016). Open-source platforms (ELAN, Praat, LaBB-CAT) and corpus management systems can be used for the creation and management of the corpus.

The most valuable material in the study of dialects is living speech (Himmelmann, 1998). Systematic field research should be conducted in various parts of the region, with high-quality audio and video recordings collected, transcribed, and turned into archives that are open while personal data is protected. These archives will serve as a primary data source not only for linguists but also for the training of AI models. The absence of sufficiently annotated and balanced databases for the Ganja-Gazakh-Karabakh dialects seriously limits the functionality of AI models. For this reason, the creation of dialect resources should be evaluated not merely as a technical project but as a strategic component of national language policy and cultural heritage preservation.

**Institutional cooperation model**

The effective application of digital language technologies is not possible through individual and scattered initiatives. Achieving sustainable results requires systematic institutional cooperation among state structures, academic scientific institutions, and the technology sector. At the state level, the Ministry of Culture, the Ministry of Education, and relevant state agencies should prepare a unified strategy and regulatory framework for the digital preservation of the language and ensure the financing of projects.

In the scientific sphere, the Institute of Linguistics of ANAS, the Nakhchivan Branch of ANAS, and the philology faculties of leading universities act as the principal expert force in the areas of scientific methodology, data verification, and analysis (Institute of Linguistics of ANAS, 2021). In the technology sector, local and international IT companies and startups can participate in the development of technological solutions, platforms, and user interfaces. To institutionalize this tripartite cooperation, two concrete structures are proposed: a National Center for Digital Language Technologies, as an entity ensuring overall coordination and the creation of resources in accordance with unified standards; and Joint Linguistics and AI Laboratories, envisaged as a structure that transforms academic knowledge into practical technological applications. In these laboratories, for example, the development of transformer-architecture language models (BERT-type variants) adapted to the characteristics of the Azerbaijani language could be carried out.

**From archiving to active use: AI applications**

Traditional approaches in the field of dialect preservation have mainly been limited to documentation and archiving. Modern digital strategies, on the other hand, bring to the fore the active use, learning, and integration of the Ganja-Gazakh-Karabakh dialects into the everyday communication environment. AI technologies create broad opportunities in this direction (Jurafsky & Martin, 2023).

Automatic speech recognition (ASR) systems make it possible to convert dialectal speech into standard text; the training of these systems requires dialect audio archives, and they can be used in real-time transcription, archive search, and educational tools. The creation of virtual assistants capable of speaking with certain regional accents or dialectal features, based on text-to-speech (TTS) technologies, serves as a powerful means of preserving the sound of the dialect across generations. Dialect-affiliation identification tools accelerate dialectological research by automatically determining which specific patois of the Ganja-Gazakh-Karabakh dialect a given text or speech fragment belongs to. Machine translation systems adapted to the dialect, in turn, can, at the initial stage at the lexical level, create the possibility of translating from standard Azerbaijani into the dialect or vice versa, thereby serving the creation of regional variants of literary texts.

In the direction of education and public outreach, gamified learning applications facilitate the learning and reinforcement of dialect words and expressions. An interactive platform called the “Dialect Map” could give users the opportunity to upload the language features of their own regions in audio or video format and become acquainted with the language materials of other regions. Virtual and augmented reality (VR/AR) applications create experiences that take the user into a virtual environment where the dialect is actively used – for example, a virtual village square. The development of lightweight applications capable of working in offline mode for mobile devices increases the accessibility of dialect resources, especially in the field of education, and strengthens their functional value.

**Standardization, digital norms, and ethical challenges**

Digital transformation also sharpens contradictions related to the standardization of the language. Spelling and grammar-checking systems, terminology platforms, and digital mechanisms applied in official document circulation strengthen the position of the standard language. This, in turn, can lead to

a narrowing of the sphere of use of the Ganja-Gazakh-Karabakh dialects (Bender, 2019). AI models tend to evaluate the variant with the most data – the standard language – as "normal" and the others as an "exception"; this can automatically push dialects into a marginal position. At the same time, the use of the standard language in official and business environments can be associated with "digital prestige", while the use of dialects can sometimes be associated with an unprofessional image (Trudgill, 2000).

Such a situation necessitates ethical and practical discussion. Language technologies should not only strengthen the normative language but should also support inclusive approaches that take dialect diversity into account (Bender, 2019). It is proposed to create separate, "alternative" rather than "corrective" models for dialects: systems should be able to offer the user a choice between the standard variant and the regional variant. The principle of user-centered design requires that technologies respect the user's language choice and not restrict it. The principle of data justice, in turn, requires attention to be paid to collecting samples from all social and age groups of the region in the creation of dialect databases. Inclusive AI models can contribute to linguistic equality in the digital environment by ensuring the parallel processing of standard and non-standard language forms.

**Lexical-semantic and toponymic aspects**

The lexical composition of the Ganja-Gazakh-Karabakh dialect group reflects the historical-geographic characteristics of the region. Museyibov's (2010) research in the field of toponymy shows that the place names of the region are an important source in the formation of the lexical layers of the dialect, and the digital registration of these names should be an inseparable part of the dialect corpus. The AI-based processing of toponyms – for example, the automatic recognition and mapping of geographic names – can serve to functionally enrich the "Dialect Map" platform. Given the performative nature of folklore genres (Bauman & Briggs, 1990), the folklore part of the corpus should be formed not only as text but also in audio and video formats.

The systematization of differences between the dialect and the standard language at the lexical level has a direct impact on the quality of machine translation models. From this point of view, the creation of dialect-standard lexical correspondence dictionaries (parallel dictionaries) should be evaluated as a priority task. Such dictionaries can serve as a primary resource both for the training of translation models and for the preparation of educational materials. The deep structure–surface structure distinction put forward by Chomsky (1965) can theoretically be considered as a methodological framework for the development of lexical-grammatical translation models between dialects, although modern transformer-based models rely more on a statistical approach.

**Integration strategy in the education system**

The sustainability of the digital transformation of dialect resources is, first and foremost, conditioned by their integration into the education system. The teaching of dialect materials within elective courses at the secondary school level can be organically combined with the study of the region's history and cultural heritage. In this regard, digital tools – interactive dictionaries, audio games, animated folklore materials – can play an effective role in engaging students' interest. At the higher education level, the development of joint curricula between philology and computer science faculties creates the basis for the formation of a "Digital Dialectology" direction.

Teacher training is also an inseparable part of this process: professional development programs should be organized to prepare teaching staff capable of using dialect materials effectively while also being able to apply basic digital tools. In the media sphere, the use of dialect materials by regional broadcasting organizations, the preparation of folklore programs in podcast format, and the dissemination of dialect content on social media platforms can help direct the attention of broad segments of the population to this topic. In the field of public services, in the long-term perspective, the development of dialect-sensitive interfaces for regional users could increase digital accessibility, although at the present stage this should be evaluated more as a prospective direction.

The implementation of this multifaceted strategy should not be limited to the availability of technological solutions alone but should also envisage the positive shaping of society's attitude toward dialects. In this regard, public awareness campaigns, the linking of regional cultural festivals with digital platforms, and the active engagement of local communities in the process are additional factors ensuring the success of the strategy.

**Risks and limitations**

A number of risks must be taken into account in the application of the proposed model. First, limited financial resources may threaten the sustainability of long-term projects; for this reason, it is expedient to attract international grant programs and scientific foundations alongside state financing. Second, the protection of the personal data of dialect speakers requires particular sensitivity; from this point of view, the anonymization of data and the documentation of user consent should be applied as mandatory procedures.

Third, the quality of technological solutions is directly dependent on the size of the database; models built on small corpora may have a high error rate, which can reduce users' trust in the technology. To minimize this risk, a phased development strategy – pilot projects at the initial stage, followed by gradual expansion – is recommended. Fourth, the balance between standardization and diversity requires continuous monitoring at the institutional level; otherwise, even systems built with good intentions may lead to unequal outcomes.

**Summary of the results**

The research conducted has shown that the three main dimensions of the digital transformation of the Ganja-Gazakh-Karabakh dialects – the resource base, institutional cooperation, and the ethical framework – are interdependent. First, the absence of annotated corpora enriched with sociolinguistic metadata is the main obstacle limiting the functionality of AI models in the dialect environment. Second, without institutionalized cooperation mechanisms among the state, science, and technology sectors, scattered projects do not produce sustainable results. Third, when standardization processes are carried out without control, AI technologies carry the risk of squeezing out dialect diversity, which makes design approaches based on the principle of alternativity necessary.

These findings are consistent with the conclusion of Grenoble and Whaley (2006) that technology is effective only as part of a broader social and educational strategy, and with the position of Bender (2019) that language technologies must make the language variants they cover explicit. Unlike the earlier works, which treat the resource, institutional, and ethical aspects separately, the proposed model links them in the material of a single regional dialect group.

As a practical outcome of the study, three priority directions have been identified: the design of the dialect corpus and the launch of a pilot collection campaign; the development and open sharing of basic open-source ASR and TTS models; and the organization of a joint training program called "Digital Dialectology" for linguists and computer scientists. The consistent implementation of these directions can secure the future of the Ganja-Gazakh-Karabakh dialects both as an object of scientific research and as a living means of communication.

**Conclusion**

The aim of the study – to develop a conceptual model reconciling the preservation of dialectal diversity with the development of the national digital language infrastructure – has been achieved by means of content analysis, comparative analysis of international experience, systemic modeling, and theoretical generalization. The digital transformation of the Azerbaijani language is a multifaceted process encompassing technological, institutional, and cultural dimensions. Although AI creates unparalleled opportunities for the preservation and revitalization of the Ganja-Gazakh-Karabakh dialects, the realization of these opportunities depends on a strong digital resource base, institutional cooperation, and ethically grounded design approaches.

The article has substantiated that inclusive AI-based language technologies present a sustainable model reconciling the preservation of dialect diversity with the development of the national digital language infrastructure. The digital transformation of the Ganja-Gazakh-Karabakh dialects and patois cannot be evaluated as a purely technical project, nor as a purely humanitarian one; it is a complex national project requiring the synthesis of both fields. Its success depends on a change in society's attitude toward dialect diversity, the mobilization of scientific-technological potential, and the sustainability of institutional will. The results can be applied in practice by state institutions, academic establishments, and technology companies through the three priority directions identified above.

Future research should be deepened in the following directions: the expansion of dialect databases and their alignment with international standards (TEI, OLAC); the improvement of AI models sensitive to regional characteristics through few-shot and transfer learning methods; the study of ways to effectively integrate these technologies into education, media, and public communication spheres;

and research into the socio-psychological dynamics of the acceptance of dialects in the digital environment. In this way, technology can become a means not only of archiving the past but also of shaping the future of the regional varieties of the language, passing them on to future generations as a living means of communication and as an unparalleled cultural heritage.

**References**

1. Akhundov, A. A. (1984). Azərbaycan dilinin dialektologiyası [Dialectology of the Azerbaijani language]. Baku: Elm. (In Azerb.)

2. Alizadeh, A. (2016). Qərbi Azərbaycan dialektləri [Dialects of Western Azerbaijan]. Baku: Elm və Təhsil. (In Azerb.)

3. Azerbaijan National Academy of Sciences Institute of Linguistics. (2021). Azərbaycan milli korpusu: perspektivlər və çağırışlar [The Azerbaijani national corpus: perspectives and challenges]. Baku: ANAS Institute of Linguistics. (In Azerb.)

4. Bauman, R., & Briggs, C. L. (1990). Poetics and performance as critical perspectives on language and social life. Annual Review of Anthropology, 19, 59–88.

5. Bender, E. M. (2019). The BenderRule: on naming the languages we study and why it matters. The Gradient. Retrieved from https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/

6. Bucholtz, M., & Hall, K. (2005). Identity and interaction: a sociocultural linguistic approach. Discourse Studies, 7(4–5), 585–614.

7. Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MIT Press.

8. Crystal, D. (2000). Language death. Cambridge: Cambridge University Press.

9. Grenoble, L. A., & Whaley, L. J. (2006). Saving languages: an introduction to language revitalization. Cambridge: Cambridge University Press.

10. Hamzayev, C. (1992). Azərbaycan dilinin dialektləri və şivələri [Dialects and patois of the Azerbaijani language]. Baku: Maarif. (In Azerb.)

11. Himmelmann, N. P. (1998). Documentary and descriptive linguistics. Linguistics, 36(1), 161–195.

12. Hinton, L., Huss, L., & Roche, G. (Eds.). (2018). The Routledge handbook of language revitalization. New York: Routledge.

13. Jurafsky, D., & Martin, J. H. (2023). Speech and language processing (3rd ed. draft). Retrieved from https://web.stanford.edu/~jurafsky/slp3/

14. Labov, W. (2001). Principles of linguistic change: social factors. Oxford: Blackwell.

15. McEnery, T., & Hardie, A. (2012). Corpus linguistics: method, theory and practice. Cambridge: Cambridge University Press.

16. Museyibov, M. (2010). Azərbaycan toponimikası [Azerbaijani toponymy]. Baku: Maarif. (In Azerb.)

17. Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational sociolinguistics: a survey. Computational Linguistics, 42(3), 537–593.

18. Shiraliyev, M. (2008). Azərbaycan dialektologiyasının əsasları [Fundamentals of Azerbaijani dialectology]. Baku: Şərq-Qərb. (In Azerb.)

19. Trudgill, P. (2000). Sociolinguistics: an introduction to language and society (4th ed.). London: Penguin Books.

20. UNESCO. (2010). Atlas of the world's languages in danger (3rd ed.). Paris: UNESCO Publishing.