T Ü R K O L O G İ Y A

№ 1 **2025**

https://doi.org/10.59849/2313-5204.2025.1.32

DİLÇİLİK
ЯЗЫКОЗНАНИЕ
LINGUISTICS

NADIR MAMMADLI (Azerbaijan)[*]
MASUD MAHMUDOV(Azerbaijan)[**]
ILHAM TAHIROV (Azerbaijan)[***]

NATIONAL CORPUS AND LEXICAL INFRASTRUCTURE: THE EXPERIENCE OF BUILDING THE AZERBAIJANI LANGUAGE LEXICAL DATABASE

Abstract

This article presents the methodology and outcomes of a multi-stage research project aimed at building the first comprehensive lexical database of the Azerbaijani language, developed within the framework of corpus linguistics and statistical lexicography. The database functions as a foundational linguistic module within the national corpus, providing structured lexical inventories for both academic research and technological applications. During the research, a corpus of 520 million word-forms from various functional styles was collected, cleaned, and structured using specialized software. As a result of automatic processing, 2, 918, 910 word-forms were initially identified; through several subsequent stages of technical and lexical filtering, the database was refined. Ultimately, 175, 521 lexical units were compiled into structured word lists, sorted by frequency and in alphabetical order. Additionally, 93, 287 words not found in the orthographic dictionary of Azerbaijani language were identified and presented in a separate list. The article elaborates on the structure of the national corpus, its components (fiction, publisistic, scientific, official, spoken, and educational texts), concordances, and linguistic analyzers. The lexical database is presented as a strategic resource for both theoretical research and practical applications such as natural language processing, automatic translation, text-to-speech systems, and artificial intelligence models. The study demonstrates significant scientific and practical outcomes in the development of digital lexical resources for the Azerbaijani language.

Keywords: *artificial intelligence, corpus linguistics, national corpus, lexical database, natural language processing, statistical lexicography, concordance*

[*] Director of the Institute of Linguistics named after Nasimi of Azerbaijan National Academy of Sciences, Prof.Dr. E-mail: nurlan1959@gmail.com

[**] Head of departament at the Institute of Linguistics named after Nasimi of Azerbaijan National Academy of Sciences, Prof. Dr. E-mail: mmasud@bk.ru

[***] Head of departament at the Institute of Linguistics named after Nasimi of Azerbaijan National Academy of Sciences, Prof. Dr. E-mail: ilham_tahir@rambler.ru

Background and Justification

By the decision of the Azerbaijan National Academy of Sciences (ANAS) dated October 21, 2024, "The Development Concept and Roadmap of ANAS for 2025-2030" was officially approved. The main goal of this document is to ensure the development of national science in accordance with modern challenges. The integration of artificial intelligence, digital and smart technologies into the humanities and social sciences has been identified as one of the priority directions in the Concept.

Within this context, various institutes of ANAS, including the Nasimi Institute of Linguistics, have been assigned a number of strategic tasks. These tasks include ensuring the representation of the Azerbaijani language in digital environments, adapting it to technological applications, developing various dictionaries and corpora, and conducting AI-based linguistic research.

In response to these challenges, the Nasimi Institute of Linguistics is currently active in the following areas:

1. Digitization and Technological Adaptation:
- Expanding the presence of the Azerbaijani language in the digital space and increasing its accessibility and usability;
- Developing electronic dictionaries of the Azerbaijani language and improving the digital versions of existing ones;
- Creating online textbooks and technologies for the teaching of the Azerbaijani language.

2. Development of Research Resources:
- Compiling concordances based on the works of Azerbaijani writers and poets, as well as written monuments of the language;
- Creating various sections of the Azerbaijani language national corpus of the and launching its initial version for online access;
- Expanding research on language learning within the context of artificial intelligence.

3. Personnel Training and International Cooperation:
- Training new scientific personnel in emerging fields of linguistics;
- Implementing joint scientific projects with foreign and local researchers, as well as strengthening international cooperation with national higher education institutions.

Within this complex framework of initiatives, the creation of a lexical database for the Azerbaijani language is of particular importance. The primary goal is to systematically collect the lexical resources of the language and make them available for both scientific research and technological fields.

The process of building the lexical database is closely connected with the development of the Azerbaijani Language National Corpus and concordances. Corpus linguistics, as a modern field of linguistics, involves collecting and analyzing texts in electronic form based on authentic language data. The national corpus encompasses various genres, styles, dialects, and other features of the

Azerbaijani language. Such corpora are important tools that provide researchers with operational and reliable linguistic information.

Unlike traditional linguistics, the corpus-based approach enables us to study language in the context of its actual use. Through search engines and analysis tools, researchers can quickly and accurately identify any lexical or grammatical unit. The effectiveness of the corpus depends on the diversity and scale of the texts it covers.

There are internationally recognized examples in this field, such as:

- *British National Corpus (BNC)* – consists of both written and oral texts, covers approximately 100 million word-forms;
- *American National Corpus (ANC)* – containing 22 million word-forms and representing various types of written and spoken language samples;
- *Turkish National Corpus (Türkçe Ulusal Derlem, TUD)* – A corpus prepared in Turkiye, consisting of 50 million word-forms, covering 39 different genres (scientific articles, novels, letters, blogs, etc.), supported by TÜBİTAK.
- Significant work has been carried out in recent years regarding the Azerbaijani Language National Corpus (Mahmudov, Fatullayev, Fatullayev, Abbasov, Abdullayev, 2016: 15-28) . Statistical-linguistic research based on samples of modern literary language, classical literature, and written monuments has laid a solid foundation in this area. The compilation of frequency and reverse alphabetical dictionaries is also an important part of this work (Vəliyeva, Mahmudov, Pines, Rahmanov, Sultanov, 1999:248).

The integration of the lexical database, corpus, and concordances contributes to the formation of a unified linguistic resource for the Azerbaijani language. While concordances reveal the vocabulary and stylistic features of specific authors or sources, the lexical database reflects frequency indicators and the general lexical landscape across functional styles. These resources provide diverse opportunities for theoretical linguistics, translation systems, language teaching, and artificial intelligence applications.

Concordance Projects

At the Department of Artificial Intelligence and Computational Linguistics of the Nasimi Institute of Linguistics (ANAS), comprehensive concordances have been developed based on the works of prominent Azerbaijani poets such as Huseyn Javid, Mikayil Mushfig, and Sabir Rustamkhanli, as well as the monumental text of the Azerbaijani language – *Kitabi-Dede Gorgud* (https://korpus.azerbaycandili.az/concordance). According to the concordance data, the total word count is as follows: 186,864 in the five-volume selected works of Huseyn Javid (Vol. I – 30,356; Vol. II – 44,284; Vol. III – 42,558; Vol. IV – 19,680; Vol. V – 49,986), 77,106 in the works of Mikayil Mushfig, 31,902 in *Kitabi-Dede Gorgud*, and 1,777,566 in the 15-volume collected works of Sabir Rustamkhanli (Vol. I– 34583, Vol. II – 36647, Vol. III – 37136, Vol. IV – 61005, Vol. V – 137369, Vol. VI – 140240, Vol. VII – 161690, Vol. VIII – 160443, Vol.

IX – 149281, Vol. X – 157171, Vol. XI – 174290, Vol. XII – 142615, Vol. XIII – 154080, Vol. XIV – 149840, Vol. XV – 71176).

Building on the experience gained in the preparation of concordances within Azerbaijani linguistics, ongoing and future efforts aim to advance this line of research by developing new concordances that encompass the language of written historical monuments, as well as the literary works of both classical and contemporary Azerbaijani authors. In the coming years, concordances will be compiled based on selected works by prominent representatives of Azerbaijani literature, including Seyid Azim Shirvani, Gasim bey Zakir, Mirza Alakbar Sabir, Mir Jalal, Rasul Rza, among others.

Among the most noteworthy concordance initiatives is the development of a concordance for the Azerbaijani translation (from Arabic) of the renowned all-Turkic written monument "Diwan Lughat al-Turk" by Mahmud Kashgari. This translation – discovered, edited, annotated, and beautifully published by Prof. Dr. Nadir Mammadli – is expected to serve as a valuable resource for further research in Turkic lexicography and historical linguistics.

In recent years, significant progress has also been made in the development of linguistic analyzers alongside technologies such as artificial intelligence, automatic text processing, text-to-speech systems, and integrated electronic dictionary corpora. The expansion of research in this direction is of particular importance.

Since the scientific groundwork has already been laid for the creation of a national corpus of the Azerbaijani language, it is now possible to launch the project on a broader scale. For this purpose, there is a need for large-scale textual material encompassing diverse domains. The reliability of a corpus depends strongly on its volume and balanced structure.

The Structural Model of the Azerbaijani Language National Corpus

In light of existing efforts and foundational work in corpus development, the establishment of a comprehensive, systematized, and balanced textual database is deemed essential for the creation of the National Corpus of the Azerbaijani Language. Accordingly, the corpus structure is proposed to cover the following components (Mahmudov, 2013: 356)

1. Main Corpus

This component comprises various functional styles and genres, including:

– *Fiction and Literary texts*: encompassing both poetry and prose, dramatic works, folklore, and samples from classical and contemporary literature;

– *Publicistic discourse*: print and digital media materials such as newspaper articles, blogs, and online publications;

– *Official and administrative texts*: legal documents, presidential decrees, speeches, and other official records;

– *Scientific and technical texts*: publications in the fields of humanities, social sciences, natural sciences, and applied sciences;

– *Religious and philosophical texts*: written materials and texts related to various religions and ideological trends;

– *Dialectological materials*: texts and transcripts that reflect the regional and dialectal variations of Azerbaijani.

2. Subcorpus of Electronic Dictionaries

This subcorpus incorporates digitized versions of various lexicographic resources in the Azerbaijani language, including synonymic, antonymic, homonymic, phraseological, terminological, dialectal, bilingual and multilingual, encyclopedic, explanatory, orthoepic, statistical, frequency-based, reverse alphabetical, and orthographic dictionaries.

3. Subcorpus of Oral Texts

Designed to capture the spoken dimension of the language, this component includes:

a) Samples of everyday and colloquial speech;

b) Public speeches and recorded interviews;

c) Audio and transcribed materials representing dialectal diversity.

4. Parallel Corpora

This section enables cross-linguistic analysis and includes:

a) *Bilingual corpora*: Azerbaijani–English, Azerbaijani–Russian, Azerbaijani–German, etc.;

b) *Multilingual corpora*: multiple translations of the same texts across different languages;

c) *Comparative translation sets*: diverse translation versions intended for contrastive studies.

5. Educational Subcorpus

This component compiles language materials designed for pedagogical purposes, such as:

a) School and university textbooks;

b) Instructional manuals, teaching materials, teaching aids, and examination tests, etc.

6. Subcorpus of Concordances

This subcorpus includes concordances constructed on the basis of:

a) The works of Azerbaijani literary figures and classical authors;

b) Written monuments and historical texts of the Azerbaijani language.

These concordances provide valuable data for semantic and stylistic analysis of lexical units.

7. Subcorpus of Linguistic Analyzers and Software Infrastructure

This includes the development and integration of:

a) Analytical modules at morphological, syntactic, semantic, and lexical levels;

b) Speech synthesis and recognition technologies;

c) Corpus access systems equipped with advanced search engines and user interfaces.

The proposed structure of the Azerbaijani Language National Corpus is intended to serve not only academic and linguistic research purposes but also practical applications in artificial intelligence, natural language processing, language education technologies, and machine translation systems.

The development of the lexical database of Azerbaijani language

The development of a comprehensive lexical database of Azerbaijani language is regarded as a primary component of the Azerbaijani Language National Corpus and is carried out through the synthesis of statistical lexicography, corpus linguistics, and natural language processing. The primary objective of this initiative is to build a representative lexical inventory derived from frequency-based and lexical analysis of a broad range of texts encompassing diverse functional styles and genres of the Azerbaijani language. To facilitate this, dedicated software has been designed for the systematic collection and processing of texts from various stylistic domains of the language.

Text Collection and Initial Processing. To ensure lexical richness, genre diversity, and functional coverage in the lexical database of the Azerbaijani language, a wide range of digital resources were utilized. These sources reflect official, institutional, cultural, and media discourse and provide representative material from different domains of modern Azerbaijani usage. The selected sources contribute to the balanced structure of the corpus by including legal-administrative registers, public communication, cultural production, and journalistic texts.

In this stage of work, written texts covering various functional styles were collected from the following publicly accessible websites:

1. Presidential Administration and Official Government Publications

https://president.az – Official website of the President of the Republic of Azerbaijan;

https://xalqqazeti.az/az–"Xalq" newspaper (official government publication);

https://www.azerbaijan-news.az/az – "Azərbaycan" newspaper (official organ of the Parliament);

https://e-qanun.az – Legislative database of the Ministry of Justice;

2. Websites of Official State Institutions (Ministries and Committees)

https://arx.com.az – State Committee for Urban Planning and Architecture;

https://culture.gov.az – Ministry of Culture;

https://eco.gov.az – Ministry of Ecology and Natural Resources;

https://justice.gov.az – Ministry of Justice;

https://fhn.gov.az – Ministry of Emergency Situations;

https://minenergy.gov.az – Ministry of Energy;

https://mfa.gov.az – Ministry of Foreign Affairs;

https://mod.gov.az – Ministry of Defense;

https://njustice.gov.az – Ministry of Justice of the Nakhchivan Autonomous Republic;

https://scara.gov.az–State Committee on Affairs with Religious Associations;

https://science.gov.az – Azerbaijan National Academy of Sciences;

https://sosial.gov.az – Ministry of Labor and Social Protection of Population;

https://stat.gov.az – State Statistical Committee.

3. Cultural, Diaspora, and Public Organizations

https://www.millikitabxana.az – Azerbaijan National Library;

https://medeniyyet.az – Cultural and artistic news portal;

https://diaspor.az – Portal on diaspora activities and Azerbaijanis abroad;

http://azyb.az – Union of Azerbaijani Writers.

4. News and Information Agencies

https://apa.az – APA (Azerbaijan Press Agency);

https://azertag.az – AZERTAC (State News Agency of Azerbaijan);

https://report.az – Report Information Agency.

5. News portals:

https://aznews.az – AzNews news portal;

https://bakupost.az – BakuPost news portal;

https://lent.az –News and information portal;

https://telejurnal.az – Media and journalism news portal.

6. Newspapers and Magazines (print or online media outlets):

https://525.az – "525-ci qəzet";

https://baki-xeber.com – Bakı-Xəbər newspaper;

https://science.gov.az/az/forms/archive/3873 – "Elm" newspaper;

https://edebiyyatqazeti.az – "Ədəbiyyat" newspaper;

https://adalet.az/az – "Ədalət" newspaper;

https://hurriyyet.az/az – "Hürriyyət" newspaper;

https://ikisahil.az – "İki Sahil" newspaper;

https://musavat.com – Online socio-political newspaper;

https://respublika-news.az – "Respublika" newspaper;

https://sesqazeti.az – "Səs" newspaper;

https://www.yeniazerbaycan.com – "Yeni Azərbaycan" newspaper,

"Azərbaycan" magazine, "Ulduz" magazine, "Qobustan" magazine.

7. Fiction and scholarly heritage:

Texts by Azerbaijani writers, poets, and scholars were collected in HTML, PDF, and DOC formats.

Writers and poets: Anar, Afag Masud, Aslan Guliyev, Bakhtiyar Vahabzada, Jafar Jabbarly, Jalil Mammadguluzada, Chingiz Abdullayev, Elchin, Akram Aylisli, Alagha Kurchayli, Ali Amirli, Ali Karim, Ali Valiyev, Alibala Hajizada, Anvar Mammadkhanli, Aziza Jafarzada, Farman Karimzada, Fikrat Goja, Firuz Mustafa, Flora Khalilzada, Khanimana Alibeyli, Huseyn Javid, Huseyn Kurdoglu, Ilyas Afandiyev, Isa Ismailzada, Isa Huseynov, Ismail Shikhli, Gilman Ilkin,

Kamal Abdulla, Karim Dunyamali, Mahira Naghiqizi, Magsud Ibrahimbeyov, Mehdi Huseyn, Madina Gulgun, Mammad Aslan, Mammad Ismail, Mikayil Mushfig, Mir Jalal, Mirza Ibrahimov, Mirza Fatali Akhundzada, Movlud Suleymanli, Musa Yagub, Nabi Khazri, Nariman Hasanzada, Nigar Rafibeyli, Nusrat Kasamanli, Nuraddin Adiloglu, Ramiz Rovshan, Rasul Rza, Rustam Behrudi, Rustam Ibrahimbeyov, Sabir Ahmadli, Sabir Rustamkhanli, Sabit Rahman, Salam Gadirzada, Seyran Sakhavat, Samad Vurghun, Sohrab Tahir, Suleyman Rahimov, Suleyman Rustam, Tofig Bayram, Vagif Samadoglu, Yusif Samadoglu, Zalimkhan Yagub.

Scholars: Aghamusa Akhundov, Bakir Nabiyev, Aziz Mirahmadov, Isa Habibbeyli, Gazanfar Kazimov, Gazanfar Pashayev, Gulu Khalilov, Mir Jalal, Nadir Mammadli, Nasiman Yagublu, Nizami Jafarov, Shirindil Alishanov, Tehran Alishanoghlu, Yagub Mahmudov, Yashar Garayev, Yusif Seyidov, Ziya Bunyadov.

The comprehensive multi-volume collections *Our Independence is Eternal* (46 volumes) by National Leader Heydar Aliyev and *Development is Our Goal* (125 volumes) by President Ilham Aliyev have been included into the text corpus as key sources reflecting political and ideological discourse in modern Azerbaijani.

Processing Stage via Software Tools. Following the compilation of the text corpus, the next stage involved the automated linguistic and technical refinement of the lexical database through dedicated software tools. The following key operations were implemented:

- *Word-form separation and frequency ranking:* A frequency list was generated based on 2,918,910 word forms extracted from the corpus.
- *Removal of misspellings and technical residuals:* Approximately 126,000 erroneous or misspellings and technically flawed word forms were detected and removed.
- *Exclusion of onomastic units from the database through structured appendices:* In order to preserve the general lexical balance of the corpus, proper names – including toponyms and anthroponyms – were separated and removed from the main dataset as follows:
  - Appendix 1 – Azerbaijani geographical names: 46,346 word forms
  - Appendix 2 – International geographical names: 13,878 word forms
  - Appendix 3 – Azerbaijani personal names: 24,009 word forms
  - Appendix 4 – International personal names: 5,392 word forms
- *Automatic filtering of numerical data and symbols:* To neutralize the influence of statistical and non-lexical elements, 22,299,902 numerical tokens and 2,479,123 alphanumeric codes were automatically removed.
- *Normalization and integration of hyphenated forms:* Word forms written with hyphens were handled specifically, and the following additions were made:
  - Hyphen at the beginning: +5,122 word forms

- Hyphen at the end: +14,014 word forms

As a result of these filtering and normalization operations, the final lexical database comprised 2,915,817 word forms.

Root and stem segmentation. In this stage, root and stem segmentation of word forms was carried out to the extent permitted by the processing software. Forms that could not be segmented automatically were analyzed manually. The conversion of word forms into lexemes was continued and refined, and updated frequency values were recalculated.

Following manual filtering and corrections, the statistical profile of the lexical database was obtained (see: Figure 1). The total number of word forms in the processed database amounted to 2,171,260. This number encompasses both lexical units and non-lexical items, such as foreign borrowings, errors, abbreviations, etc.

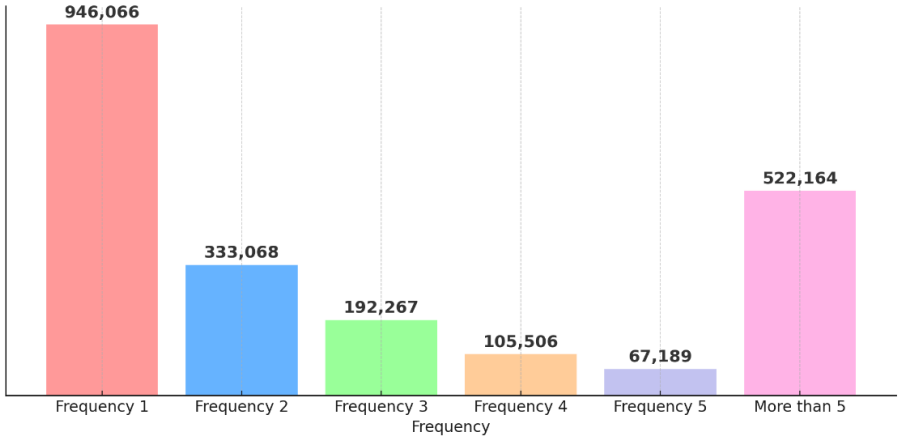

Figure 1. Distribution of Word Frequencies

Word forms that occur only once (Frequency 1) represent approximately 43.6% of the entire lexical database. This is a typical phenomenon in corpus linguistics, frequently reflecting either rare lexical items or noise factors (errors) such as misspellings, typographic errors, and transliteration inconsistencies or differences.

Lexical inits with frequencies greater than 5 constitute over 24.03% of the dataset and predominantly reflect core functional and general vocabulary.

Final Cleaning and Lexicographic Structuring. In the final processing stage, the database underwent an extended phase of technical and linguistic refinement**.** The principal aim at this point was the elimination of non-standard, technically induced, and ultra-low-frequency items**,** thereby facilitating the transition of the word database into a structured form suited for lexicographic applications and linguistic research.

At the final stage of lexical filtering, word-forms were systematically removed from the database based on their frequency distribution (See: Figure 2):

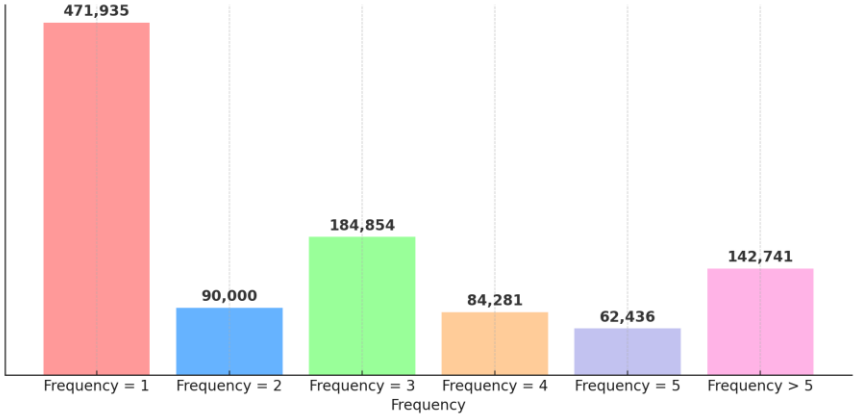

Figure 2. Distribution of deleted items by frequency

As a result, a total of 1,036,247 word-forms were excluded from the lexical database.

Wordlist Compilation. Upon completion of the cleaning and analysis procedures, four main wordlists were compiled, each serving specific linguistic and lexicographic functions:

1. Frequency-ordered wordlist (Frequency List-I):

– Comprises 175,521 entries;

– Items are sorted in descending order of frequency;

– Based on this list, it is possible to identify and extract high-frequency words in the Azerbaijani language.

2. Alphabetically ordered wordlist (Alphabetical List-II):

– Contains the same 175,521 items;

– Words are arranged in alphabetical order, with their corresponding frequency values indicated alongside each entry. To determine the frequency of a particular word, one must locate it in the alphabetical list and refer to the associated frequency indicator.

3. Frequency list of out-of-dictionary entries (Frequency List-III):

– Includes 93,287 items

- This list presents the frequency distribution of words that are included in the lexical database but not found in the *Orthographic Dictionary of the Azerbaijani Language* (Baku: Elm, 2021);

4. Alphabetical list of out-of-dictionary entries (Alphabetical List-IV):

– Also comprises 93,287 items;

– Entries are alphabetically sorted and supplemented with frequency indicators.

The third and fourth lists (Frequency List-III and Alphabetical List-IV) constitute a valuable reference for the development of future editions of orthographic and explanatory dictionaries of the Azerbaijani language.

The creation of the Azerbaijani language lexical database is not merely a lexicographic initiative; it constitutes a strategic project that lays the groundwork for the digital transformation of the Azerbaijani language. Based on this resource, it is possible to construct linguistic corpora, dictionaries, educational tools, and language technologies that will contribute to the scientific and technological development of Azerbaijani. This database enables the identification of the scope of functional vocabulary and the establishment of word inventories across various frequency levels. These capabilities are visually illustrated in Figure 3.

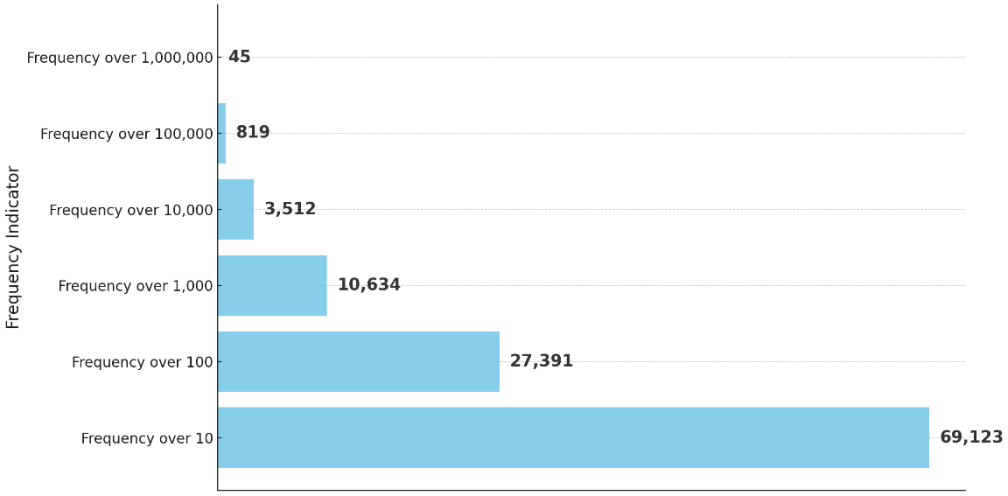

Figure 3 Word count by frequency indicator

The project for developing the Azerbaijani language lexical database was implemented under the supervision of the President of ANAS, Academician Isa Habibbeyli, and the Director of the Nasimi Institute of Linguistics, Prof. Dr. Nadir Mammadli.

The collection, classification, and both manual and automated linguistic processing of lexical materials – including step-by-step cleaning and recommendations for software-assisted cleaning – was carried out by Prof. Dr. Masud Mahmudov, Head of the Department of Artificial Intelligence and Computational Linguistics, and Prof. Dr. Ilham Tahirov, Head of the Department of Indo-European Languages, both affiliated with the Nasimi Institute of Linguistics.

Conclusion

This study presents a systematic overview of the theoretical foundations and practical stages involved in the development of the Azerbaijani language lexical database. It formulates key principles for constructing language resources through a synthesis of linguistic and computational approaches.

The lexical database can be viewed as:

a) the core structural component of the Azerbaijani National Corpus;

b) a product of both automatic and semi-automatic lexical processing methodologies;

c) a large-scale, genre-diverse linguistic resource that exceeds existing corpus-based projects in both scope and lexical coverage.

The project has resulted in the compilation of over 175,000 lexical units presented in frequency and alphabetical wordlists, and the identification of more than 93,000 lexical items absent from the current Orthographic Dictionary of the Azerbaijani Language (Baku: Elm, 2021).

Future directions include the integration of this lexical database into the national dictionary strategy, with its utilization as a reference framework for forthcoming orthographic, explanatory, terminological, bilingual, and multilingual dictionaries. Furthermore, the expansion of artificial intelligence and natural language processing technologies is essential to leverage this database for speech recognition, machine translation, chatbot development, and training of language models in Azerbaijani.

The database also holds significant potential for the development of frequency-based educational resources, the creation of lexical minimums, and the compilation of teaching materials and conversational content. In addition, forthcoming efforts should focus on extended data cleaning, refinement, and implementation of morphological tagging.

The integration of phonetic, morphological, and syntactic attributes of lexical units into the database structure is recommended as a valuable enhancement. Equally important is the development of an open-access online platform equipped with a search-enabled web interface and statistical tools, enabling researchers and the broader public to interact with and benefit from this linguistic resource.

REFERENCES

1. Statistical Analysis of the Language of “Kitabi-Dede Gorgud”. (1999). /Compilers: K.A. Vəliyeva, M.A. Mahmudov, V.Y. Pines, J.A. Rahmanov, V.S. Sultanov/. – Baku: Elm [in Azerbaijani]
2. Mahmudov, M., Fatullayev, R., Fatullayev, A., Abbasov, S., Abdullayev, N. (2016). Theoretical and Applied Issues in the Development of NLP Systems and the National Corpus for the Azerbaijani Language. – Baku: Turkologiya, No. 4 [in Azerbaijani]
3. Reverse Alphabetical Dictionary of the Azerbaijani Language (2004). /Compilers: M. Mahmudov, A. Fatullayev/. – Baku: Nurlan [in Azerbaijani]
4. Frequency Dictionary of the Azerbaijani Language (Word Roots), Volume I. (2010) /Compilers: M. Mahmudov, A. Fatullayev, et al./ – Baku: Elm [in Azerbaijani]

5. Mahmudov, M. (2017). Preconditions and Optimal Structure for the Creation of the National Corpus of the Azerbaijani Language. In: Время собирать камни …, Collection of Articles. – Baku: Mutercim [in Azerbaijani]
6. Mahmudov, M. (2013). Computational Linguistics. – Baku: Elm və Təhsil [in Azerbaijani]
7. Mahmudov, M. (2018). National Corpora of Turkic Languages. – Baku: Elm və Təhsil, – 392 p. [in Azerbaijani]
8. Mahmudov, M., Tahirov, I., Ayda-zade, K., Talibov, S. (2019). The Integrated Electronic Dictionary System as a Stage in the Creation of the Azerbaijani National Corpus. – Baku: Turkologiya, No. 1 [in Azerbaijani]
9. Mahmudov, M. (2024). Linguistic Problems of Artificial Intelligence. – Baku: Elm və Təhsil [in Azerbaijani]
10. Alphabetical-Frequency Dictionary of Məhəmməd Füzuli's Poetic Works (2004). /Compilers: K.A. Vəliyeva, M.A. Mahmudov, J.A. Rahmanov, V.S. Sultanov/. – Baku: Elm [in Azerbaijani]
11. https://korpus.azerbaycandili.az/concordance

Nadir Məmmədli (Azərbaycan)
Məsud Mahmudov (Azərbaycan)
İlham Tahirov (Azərbaycan)

Milli korpus və leksik infrastruktur: Azərbaycan dilinin söz bazasının yaradılması təcrübəsi

Xülasə

Məqalədə Azərbaycan dilinin söz bazasının hazırlanması ilə bağlı həyata keçirilən nəzəri və praktiki mərhələlər sistemli şəkildə təhlil edilmiş, dil resurslarının formalaşdırılması üçün zəruri olan texnoloji və linqvistik yanaşmalar əsasında söz bazasının qurulması prinsipləri ərsəyə gətirilmişdir. Söz bazası: a) Azərbaycan dilinin milli korpusunun əsas struktur komponentlərindən biri kimi nəzərdən keçirilmişdir; b) leksik vahidlərin mənbə əsasında avtomatik və yarıavtomatik emalı təcrübəsi əsasında yaradılmışdır; c) fərqli funksional üslubları, çoxsaylı mənbələri və böyükhəcmli materialları əhatə etməsi baxımından, mövcud korpustəməlli təşəbbüslərlə müqayisədə daha geniş miqyaslıdır. Müəlliflər hesab edirlər ki, strukturlaşdırılmış leksik inventarı təmin edən belə bir söz bazası həm akademik tədqiqatlar, həm də texnoloji tətbiqlər üçün zəruridir. Araşdırma zamanı müxtəlif funksional üslublara məxsus 520 milyon söz-forma korpusu toplanmış, xüsusi proqram təminatı ilə təmizlənmiş və strukturlaşdırılmışdır. Avtomatik emal nəticəsində ilkin mərhələdə 2.918.910 söz-forma müəyyən olunmuş, sonrakı bir neçə mərhələdə texniki və leksik filtrləmə nəticəsində baza təmizlənmişdir. Son nəticədə 175.521 leksik vahiddən ibarət tezlik və əlifba sözlükləri tərtib olunmuşdur. Eyni zamanda, Azərbaycan dilinin orfoqrafiya lüğətində yer almayan 93.287 söz müəyyənləşdirilərək ayrıca siyahı şəklində təqdim olunmuşdur. Məqalədə milli korpusun strukturu, onun bölmələri (bədii, publisistik, elmi, rəsmi, şifahi və tədris mətnləri), konkordanslar və linqvistik analizatorlar geniş şəkildə şərh olunur. Söz bazası həm nəzəri tədqiqatlar, həm də təbii dilin emalı, avtomatik tərcümə, səsləndirmə sistemləri və süni intellekt modelləri üçün

strateji resurs kimi təqdim olunur. Məqalədə Azərbaycan dilinin rəqəmsal leksik resurslarının formalaşması istiqamətində mühüm elmi və tətbiqi nəticələr ortaya qoyulur.

Açar sözlər: *süni intellekt, korpus dilçiliyi, milli korpus, söz bazası, təbii dilin emalı, statistik leksikoqrafiya, konkordans*.

Надир Мамедли (Азербаджан)
Масуд Махмудов (Азербаджан)
Ильхам Тахиров (Азербаджан)

Национальный корпус и лексическая инфраструктура: опыт создания словарной базы азербайджанского языка

Резюме

В статье представлен комплексный научно-технический анализ проекта по созданию лексической базы данных азербайджанского языка, разработанной в рамках корпусной лингвистики и статистической лексикографии. Лексическая база данных: а) рассматривается как один из основных структурных компонентов национального корпуса азербайджанского языка; б) создана на основе опыта автоматической и полуавтоматической обработки лексических единиц на основе источника; в) является более обширной, чем существующие корпусные инициативы, с точки зрения охвата различных функциональных стилей, множественных источников и больших объемов материалов. Авторы считают, что такая база данных, предоставляющая структурированные лексические инвентари, необходима как для академических исследований, так и для технологических приложений. В ходе исследования был собран, очищен и структурирован с помощью специализированного программного обеспечения корпус из 520 миллионов словоформ различных функциональных стилей. В результате автоматической обработки было первоначально выявлено 2 918 910 словоформ; после нескольких последующих этапов технической и лексической фильтрации база данных была уточнена. В конечном итоге 175 521 лексическая единица была составлена в структурированные списки слов, отсортированные по частоте употребления и в алфавитном порядке. Кроме того, были выявлены и перечислены отдельно 93 287 слов, не встречающихся в орфографическом словаре азербайджанского языка. В статье подробно рассматривается структура национального корпуса, его компоненты (художественная литература, публицистика, научная литература, официальные, устные и учебные тексты), конкордансы и лингвистические анализаторы. По мнению авторов, лексическая база данных является стратегическим ресурсом как для теоретических исследований, так и для практических приложений, таких как обработка естественного языка, автоматический перевод, системы преобразования текста в речь и модели искусственного интеллекта. Исследование демонстрирует значительные научные и практические результаты в разработке цифровых лексических ресурсов азербайджанского языка.

Ключевые слова: *искусственный интеллект, корпусная лингвистика, национальный корпус, лексическая база, обработка естественного языка, статистическая лексикография, конкорданс*.