ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul101

Problems of Artificial Intelligence for the Azerbaijani Language: The Impact of Limited Language Corpus

Raksana Aliashrafova[1]; Namiq Abdurahmanov[2]; Gunel Aghammadova[3]; Vafa Atayeva[4]

[1;2;3;4]Shaki Regional Scientific Centre, Azerbaijan National Academy of Sciences, Shaki, Azerbaijan.

Publication Date: 2026/07/17

Abstract: Artificial intelligence (AI) systems, such as large-scale speech recognition systems, are vital for acquiring linguistic competence and achieving this is essential for their use. Thus, for large-scale speech recognition systems, acquisition of linguistic competence is essential for usage of the system. Even for the widely spoken language like English, Mandarin and Spanish, these corpora are vast and have hundreds of billions of tokens, which can be used to power up virtually all natural-language processing (NLP) tasks. In the online world, however, the Azerbaijani language is far less well resourced: publicly available language corpora are limited to a small number of high-resource languages. In the article, the main problems related to corpus scarcity, such as poor quality of machine translation, incorrect morphological processing, recognition dialects of the Azerbaijan language and failure of chatbots to understand the Azerbaijan language are discussed.

Keywords: *Azerbaijani Language, Natural Language Processing, Language Corpus, Low-Resource Languages, Machine Translation, Aİ Language Models, Turkic Languages.*

How to Cite: Raksana Aliashrafova; Namiq Abdurahmanov; Gunel Aghammadova; Vafa Atayeva (2026) Problems of Artificial Intelligence for the Azerbaijani Language: The Impact of Limited Language Corpus. *International Journal of Innovative Science and Research Technology*, 11(7), 191-194. https://doi.org/10.38124/ijisrt/26jul101

I. INTRODUCTION

The development of the language competence by using Artificial Intelligence (AI) systems, in particular the big large-scale speech recognition systems is very relevant. So, for larger scale speech recognition systems the competence of the language is learnt first and then the system is used. AI overall has come a long way over the past 10 years and the advent of large language models (LLMs), trained on a wide variety of multilingual data sets has been a giant step (Hajili, N. (2024)).

The Azeri language is spoken by approximately 35 million people worldwide, and one of the least resource rich languages in the field of AI. The digital version of the Azerbaijani language is far from what is needed to create competitive language models. The Common Crawl dataset shows that the Azerbaijani language accounts for less than 0.05% of the global web corpus, which greatly limits the quality of language tools that can be used for AI applications in the Azerbaijani language.

The main research question in this article is: What are the problems faced by artificial intelligence in the absence of enough textual material for the Azerbaijani language? The paper has been organized in the following five steps: First, the theoretical background of language corpora and language models is explored; second, the existing situation of Azerbaijani digital resources is mapped; third, a set of observable errors and system failures is investigated; fourth, possible interventions based on similar approaches in related Turkic languages are examined.

A. Theoretical Background: Language Corpora and AI Language Models

A language corpus is a collection of authentic text, which is systematically prepared, stored electronically, structured and clearly defined. NLP systems today are built on the empirical base given by the corpora. Modern language models are not handcrafted with a grammar, but rather they acquire the language patterns: syntax, semantics and pragmatics, from the statistical regularities in a large set of raw text. Thus, the quality of a model can be directly limited by the size, variety and quality of the corpus on which it is trained.

The size of the corpus has no straight correlation to the performance of the model. Below the critical threshold it is found that the performance is nonlinearly reduced. This is a very stringent threshold in morphologically rich languages

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul101

since the model has to learn how to properly split and compose inflected forms. As in other languages, Azerbaijan is an agglutinative language with no separate function words but with successive suffixation containing the grammatical information.

GPT-2 and BERT have a word size greater than 3 billion words. Although it is significantly smaller in size compared to English-language models, AzCorpus is the largest publicly available corpus in the Azerbaijani language and a valuable project (Kishiyev, (2023)).

B. How Artificial Intelligence Acquires Language Knowledge

Modern language models learn the knowledge of language by a procedure known as self-supervised pre-training. In this step, the model is given billions of sentences of text and is trained to predict either the missing or following word in the sentence, a process that infers grammar, world knowledge, and discourse structure features, without requiring explicit human annotation (Hajili, N. (2024)).

While there are systematic gaps from Turkish which can be partially filled by the means of transfer learning, the lack of phonological variety and lexical innovations, as well as Russian loans in modern use creates gaps in the system which cannot be filled by the means of transfer learning alone.

C. Core AI Challenges for the Azerbaijani Language

- *Deficiencies in Machine Translation*

The overlap between the n-grams of machine translation (MT) output and those of the manually produced MT references are the basis for the measure of quality of MT, the BLEU score. Benchmark tests show a marked difference between the scores of the best MT systems in Azerbaijani and scores of similar high resource MT systems. For the translation from English to Azerbaijani, the BLEU score is 18–26 and for the English to French translation, the BLEU score of the latest neural systems is 45–55 (Salimov and Mammadova, (2023)). Furthermore these errors do not happen once in a while but they are prevailing throughout; therefore they can be termed as structurally pervasive MT errors(Salimov and Mammadova, (2023)).

Common types of errors include case suffix errors, verb aspect and tense errors, and word order errors in embedded clauses. More importantly, these are not random but systematic, reflecting persistent gaps in what the model has learned in its representations. If a construction has not been seen sufficiently in the training corpus, the model resorts to the most frequent pattern — most often Turkish or copied from the source language — and the result will be unnatural or incomprehensible to a native speaker (Karimova, S. and Pecina, P. (2023)).

- *Morphological and Grammatical Processing Errors*

The main issue is that models trained with a small size of data are forced to confront the problem of the Azerbaijani language with its morphological construction. The language has vowel harmony rules in that suffixes are chosen depending on the type of the stem's vowel. The learner must be given an enormous amount of stem-vowel-suffix combinations in various domains in order to learn the vowel harmony correctly. Although a corpus is available, most of the grammatically regular but seldom used forms are not learnt and lead to problems in text generation and parsing. (Nouri, M., Amani, M., Zohrabi, R., and Asgari, E. (2023))

For a long time, however, no reliable lemmatiser of the Azerbaijani language for use in production NLP pipelines has existed, because of the high error rates of the available tools (Jafarov and Gasimova, (2020)). Not having an annotated morphological corpus compounds the challenges, because morphological analysis by supervision cannot be done without it, and it is also impossible to check the accuracy of the obtained results.

- *Dialect Recognition and Sociolinguistic Variation*

The Azerbaijani language is not one uniform variety. The differences between North Azerbaijani spoken in the Republic of Azerbaijan and South Azerbaijani spoken in the north-western parts of Iran with respect to phonological, lexical and grammatical aspects are numerous. Inside the Republic, regional dialects have distinct characteristics that can significantly hinder automatic recognition and processing. The texts in digital corpora to date consist predominantly of conventional formal sources — mass media and official documents — with very little dialectal or colloquial style text at all. (Nouri, M., Amani, M., Zohrabi, R., and Asgari, E. (2023))

The impact of this gap in representation is immediate in speech recognition and conversational AI. Unless a voice assistant is trained to deal with informal Azerbaijani, words specific to local usage, and admixtures of Azerbaijani with Russian — which are common in urban Azerbaijani speech — it is not able to understand them. Under-representation of informal language also has repercussions for sentiment analysis and social media monitoring systems. AI systems without dialectal data will only be able to operate with a formalised version of the language, excluding extensive parts of the actual use of the language by the majority of its speakers. (Nouri, M., Amani, M., Zohrabi, R., and Asgari, E. (2023))

- *Chatbot Comprehension and Response Quality*

The use of language is not merely grammatical, but also practical: conversational AI systems — chatbots and dialogue agents — have to grasp not only grammatical constructions but also various dimensions such as context, implicature, and the customary ways of communicating in a given culture. When prompted in Azerbaijani, major commercial chatbots often switch to the processing language of Turkish or Russian and then transliterate back to Azerbaijani orthography, which leads to grammar mistakes and an overall deficiency of the output (Hajili, N. (2024)).

It was noticed during documented user tests that chatbot answers in Azerbaijani can be tense-inconsistent and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul101

semantically deviant, even including the use of Turkish lexical units in place of Azerbaijani equivalents. These errors do not necessarily relate to the architecture of modern transformer models, but rather are due to the absence of a training signal needed to calibrate Azerbaijani-specific linguistic behaviour. This creates a significant loss of functionality for Azerbaijani users of AI assistants, with tangible impacts on education, healthcare, legal services, and civic engagement. (Hajili, N. (2024)).

D. Comparative Overview: Azerbaijani vs. Selected Languages

Table 1 below presents Azerbaijani in the context of other languages, in terms of MT quality, the size of the language's respective corpus, and the availability of NLP tools, which range from well-resourced English to variously resourced Turkic languages. The data show a regularity that is manifested consistently: the size of the corpus is significantly correlated with translation accuracy and the availability of NLP infrastructure. (Sorokin, A., Shavrina, T., and Lyashevskaya, O. (2021))

Table 1 Comparative NLP Resource Profile: Selected Languages (2024 Estimates).
*AzCorpus Document Count; Token Count Estimated.

Language	Corpus Size (tokens)	MT Quality (BLEU)	NLP Tool Availability
English	~600 billion	45–55	Extensive
Turkish	~12 billion	38–44	Moderate–High
Azerbaijani	~1.9 million*	18–26	Limited
Kazakh	~3 billion	22–30	Developing
Kyrgyz	~0.8 billion	14–20	Very Limited

The documented Azerbaijani corpus is minute compared to the English corpus — it is 315,000 times smaller — which represents an enormous capability gap. Even languages as widely recognised as low-resource cases like Kazakh and Kyrgyz boast a much more powerful digital presence than Azerbaijani. The BLEU-score column shows the direct impact of these differences in resources on the actual performance of the systems in real life. (Sorokin, A., Shavrina, T., and Lyashevskaya, O. (2021))

E. Structural Causes of Digital Under-Resourcing

A combination of factors has led to under-resourcing of Azerbaijani in AI training data. Many of Azerbaijan's Soviet literary and academic works were written in the Cyrillic script, in contrast to the Latin script used by other languages, and that presents a unique challenge for their digitisation. The change to the Latin alphabet in 1991 brought additional discontinuity in that pre-transitional texts not only must be digitised but also transliterated so that they can be added to a unified NLP corpora(Kishiyev, (2023)).

In addition to these challenges, significant economic factors are at play. Large-scale development of annotated corpora is very costly and requires large long-term investments in data collection, cleaning, annotation and validation. Commercial AI language companies also tend to invest the most in the markets with the most potential users, reinforcing the self-reinforcing tendency of market marginalization of particular low-resource languages. At policy level, there is also no coordinated effort at the national level to have a systematic corpus development that can support the development of NLP competitively at the national level. (Kishiyev, H. (2023))

II. PROPOSED SOLUTIONS AND POLICY RECOMMENDATIONS

The suggestions given below were drawn from the experiences of similar programmes for corpus development of related Turkic languages, as well as from successful programmes for low-resource languages in the world.

- There is a need for a systematic digitisation programme for literature, public works and archives of Azerbaijan. The rich written tradition of Azerbaijan is represented by literature, science, law and newspapers, which are partly not digitised and accessible for NLP researchers. Such a national digitisation project — as has been carried out for the Welsh, Irish and Basque languages — would cover billions of data items and also preserve an invaluable source of cultural heritage that is at risk of falling outside the digital process.
- The materials in school and university curricula, integrated with publicly available corpora, will give high-quality and well-edited prose texts covering different subject areas. Learning materials are particularly valuable because they are written with accuracy and clarity, have wide thematic areas, and reflect the most appropriate level of written language immediately needed in the NLP toolbox. They could add substantial volume and register variation to open-access corpora that are not possible with web data alone.
- Human-centred and crowdsourced data gathering of dialects is crucial to create AI systems which can meet the needs of all Azerbaijani speakers fairly. The complete collection of recorded speech and written data from native speakers from various regions would allow for surmounting the almost complete lack of such data that exists at present. A majority of online platforms were inspired by Mozilla Common Voice and have successfully proven that such an approach is feasible for minority and resource-poor languages, creating large, diverse and openly licensed corpora with the participation of volunteers.
- There needs to be greater, more concerted investment in Azerbaijani AI research. National science foundations and international development organisations should be engaged for sharing evaluation facilities, NLP tools and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul101

corpus construction support. Relying on resources created for Turkish and Kazakh and on cooperation within the Turkic-language NLP community, it is possible to substantially increase the tempo and prevent the fruitless duplication of efforts by national programmes.

- Creating a national language corpus infrastructure with open access would give research and commercial AI development a stable backbone, serving as an environment in which the Azerbaijani language will be used with stability and openness in researches and technologies developed on this basis.

III. CONCLUSION

As shown in the present article, Azerbaijani AI systems are plagued by severe and deep-rooted issues caused by the lack of an adequate amount of digital text resources as compared to the needs of contemporary language models. The quality of machine translations is far below that of high-resource languages, morphological processing is limited, and conversational AI systems often generate sentences that are not coherent, grammatically correct, or semantically accurate. They are significant problems in making sure Azerbaijani speakers are given the same opportunities and benefits of the AI revolution.

Such structural — historical, economic and policy — underlying causes require structural responses. Advances in individual AI systems will not be enough if the underlying foundations on which advances must be built are not improved. It is both feasible and long overdue for a programme of corpus building, digitisation and NLP infrastructure to be carried out across the country over an extended period of time.

Today's investment in digital linguistic presence will see languages gaining from AI capacities today available exclusively to a handful of globally dominant languages. Those who do not invest risk further widening a linguistic digital divide, with lasting impacts for education, economic opportunity, and cultural vitality. Creating strong and capable AI for the Azerbaijani language is thus not just a technical challenge, but a question of language equity and national strategy.

REFERENCES

[1]. Hajili, N. (2024). Open foundation models for the Azerbaijani language. arXiv preprint arXiv:2407.02337.
[2]. Jafarov, Y. M., & Gasimova, R. T. (2020). Some problems in the design of domain names with the Azerbaijani alphabet on the internet. Journal of Information Society Problems, 11(1), 45–52.
[3]. Kishiyev, H. (2023). Introducing AzCorpus: A game-changing resource for NLP applications in the Azerbaijani language. Medium.
[4]. Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A robustly optimised BERT pretraining approach. arXiv preprint arXiv:1907.11692.
[5]. Nouri, M., Amani, M., Zohrabi, R., & Asgari, E. (2023). The language model, resources, and computational pipelines for the under-resourced Iranian Azerbaijani. Findings of ACL: AACL-IJCNLP 2023.
[6]. Salimov, R., & Mammadova, G. (2023). Evaluation of neural machine translation for the Azerbaijani–English language pair. Proceedings of the Workshop on Language Resources for Turkic Languages.
[7]. Sorokin, A., Shavrina, T., & Lyashevskaya, O. (2021). A large-scale study of machine translation in the Turkic languages. Proceedings of ACL 2021.
[8]. Karimova, S., & Pecina, P. (2023). Integrated approach to adapting open-source AI models for machine translation of low-resource Turkic languages. Computers, 15(2), 73.
[9]. Kartal, J., et al. (2024). Enhancing language learning through technology: A new English–Azerbaijani parallel corpus. arXiv preprint arXiv:2407.05189.