TURKIC LANGUAGES ON SEARCH ENGINE PERFORMANCE AND APPROACHES TO MITIGATION

Author: Soliyev Ulanbek, Student of Japan Digital University
Academic Supervisor: Lochinbek Boboev, PhD, Head of IT department of Japan Digital University

Abstract

This paper investigates the impact of digital content scarcity in Turkic languages on the effectiveness of modern search engines. While English dominates the global web, many Turkic languages—including Uzbek, Kazakh, Kyrgyz, and Turkmen—remain significantly underrepresented in digital content, limiting access to reliable information in native languages. The study examines the primary causes of this imbalance and discusses how limited digital resources affect search engine performance and information retrieval. Based on recent academic literature and statistical evidence, the paper argues that increasing user-generated digital content through community-driven platforms can improve information accessibility for low-resource languages. As a conceptual case study, the **Andisha** platform is presented to illustrate how multilingual infrastructure and community participation may contribute to expanding digital knowledge resources for Turkic-speaking communities. The findings highlight that strengthening digital content ecosystems is essential for improving search quality, knowledge accessibility, and the long-term digital presence of Turkic languages.

Keywords: Turkic languages, digital content, search engines, information retrieval, low-resource languages, community platforms, Andisha.

Introduction

The Internet has become one of the primary sources of information in modern society, with billions of users relying on search engines such as Google and Bing to access

knowledge on a daily basis. The effectiveness of these search engines largely depends on the availability, quality, and diversity of digital content indexed on the web.

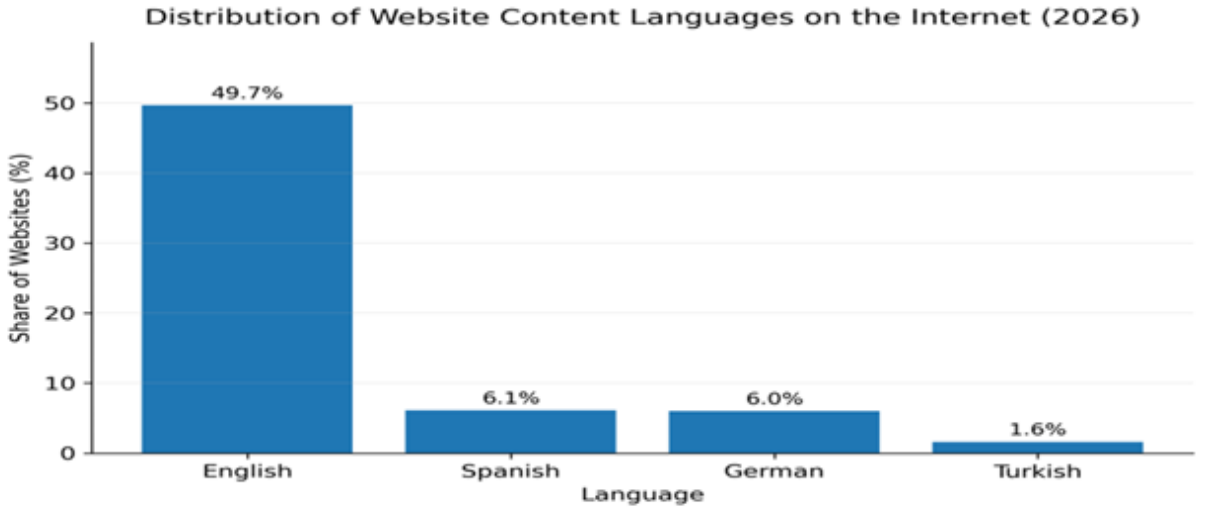
Although the volume of online information has grown rapidly over the past decades, this growth has not been distributed equally across languages. While English and several other high-resource languages dominate the global web, many Turkic languages—including Uzbek, Kazakh, Kyrgyz, and Turkmen—remain significantly underrepresented in digital content. As a result, users frequently encounter difficulties when searching for reliable information in their native languages.

The scarcity of digital content extends beyond search quality and affects broader areas such as education, scientific communication, professional development, and knowledge sharing. In technical domains, including software engineering, artificial intelligence, and information technologies, the availability of high-quality resources in Turkic languages remains particularly limited, forcing many users to rely on foreign-language materials.

This study investigates how digital content scarcity influences search engine performance in Turkic languages. It examines the underlying causes of this imbalance, reviews recent findings on information retrieval in low-resource languages and discusses community-driven approaches for increasing digital content. Furthermore, the paper presents the **Andisha** platform as a conceptual case study illustrating how multilingual community platforms may contribute to improving digital knowledge accessibility for Turkic-speaking communities.

Digital Content Scarcity in Turkic Languages

Internet language statistics clearly demonstrate the unequal distribution of digital content across languages. According to **W3Techs** (June 2026), **49.7%** of websites publish content in English, followed by Spanish (**6.1%**) and German (**6.0%**). Turkish, the largest language within the Turkic language family, accounts for only **1.6%** of websites and ranks tenth globally. [1]



Source: W3Techs, Usage Statistics and Market Share of Content Languages for Websites, 2026

The situation is considerably more severe for other Turkic languages. Uzbek, Kazakh, Kyrgyz, and Turkmen each represent **less than 0.1%** of indexed websites, placing them among the least represented languages on the web. [1] Considering that more than **200 million people** speak a Turkic language worldwide, this disparity highlights a substantial imbalance between demographic significance and digital representation. [3] The shortage of digital content is particularly evident in technical and scientific fields. Educational materials related to software engineering, artificial intelligence, cybersecurity, and other emerging technologies remain scarce in many Turkic languages. Consequently, users often depend on English-language resources to obtain accurate and up-to-date technical information.

Table 1. Comparison of Turkic Languages by Number of Speakers and Web Presence

Language	Estimated Number of Speakers	Share of Web Content
Turkish	90 million	1.6%
Uzbek	40 million	<0.1%
Kazakh	20 million	<0.1%
Kyrgyz	7 million	<0.1%

Source: W3Techs (2026), Wikipedia Turkic Languages (2026)

Note: Despite having a substantial number of speakers, Turkic languages (other than Turkish) have a disproportionately minimal presence on the web, with each accounting for less than 0.1% of web content.

Recent studies also indicate that low-resource languages receive limited support from modern language technologies. Research on Kazakh language models, for example, demonstrates that insufficient training data and linguistic resources continue to restrict the performance of large language models and other natural language processing systems. [2] These limitations indirectly affect search engines, machine translation, and AI-based information services.

Overall, the evidence suggests that the primary challenge is not the number of native speakers but the limited availability of high-quality digital content. Increasing the volume of reliable, searchable, and user-generated resources is therefore essential for strengthening the digital presence of Turkic languages and improving information accessibility.

Impact on Search Engine Performance

Modern search engines such as Google and Bing rely on machine learning–based ranking algorithms rather than simple keyword matching. The effectiveness of these algorithms depends heavily on the availability of indexed documents, user interaction data, and language-specific training resources. While such resources are abundant for high-resource languages, they remain limited for many Turkic languages, resulting in lower search quality. [4]

Recent studies confirm that search engines generally perform less effectively for low-resource languages. Queries submitted in these languages often produce incomplete, poorly translated, or less relevant results compared with equivalent searches in English, reflecting the limited availability of high-quality indexed content. [5]

The problem is further compounded by the shortage of linguistic resources required for modern information retrieval systems. Cross-lingual retrieval, multilingual language models, and semantic search technologies continue to face significant challenges when processing low-resource languages due to limited training corpora and complex linguistic structures. These limitations are particularly relevant to agglutinative Turkic

languages, whose rich morphological characteristics increase the complexity of language processing tasks. [6][7][8]

As illustrated in **Figure X**, digital content scarcity creates a self-reinforcing cycle. Limited online content reduces search quality, encouraging users to rely on foreign-language resources instead of producing content in their native languages. Consequently, the amount of searchable digital content remains insufficient, further weakening search engine performance.

These findings indicate that the primary limitation is not the search algorithms themselves, but the scarcity of high-quality digital content available for indexing. Therefore, improving search performance in Turkic languages requires expanding the volume and diversity of reliable digital resources alongside continued advances in search technologies.

Table 1. Comparison of Turkic Languages by Number of Speakers and Web Presence

Language	Estimated Number of Speakers	Share of Web Content
Turkish	90 million	1.6%
Uzbek	40 million	<0.1%
Kazakh	20 million	<0.1%
Kyrgyz	7 million	<0.1%

Source: W3Techs (2026), Wikipedia Turkic Languages (2026)

Note: Despite having a substantial number of speakers, Turkic languages (other than Turkish) have a disproportionately minimal presence on the web, with each accounting for less than 0.1% of web content.

Approaches to Mitigating Digital Content Scarcity

Improving search engine performance in Turkic languages requires expanding the availability of high-quality digital content. One of the most effective approaches identified in the literature is **community-driven content creation (crowdsourcing)**, which enables users to collaboratively produce and maintain online knowledge.

Wikipedia, available in more than **330 languages** with over **190 million articles**, demonstrates the effectiveness of this model. Community participation and network

effects enable the continuous growth of large-scale knowledge repositories. [9][10][11] However, existing global platforms do not fully address the needs of low-resource language communities. Limited language technologies, insufficient reference materials, and platform designs optimized for high-resource languages continue to hinder content creation in underrepresented languages. [9]

Based on the reviewed literature, four principles are essential for strengthening digital content development in Turkic languages:

- **Language-aware infrastructure** supporting diverse writing systems and linguistic characteristics.
- **Low barriers to participation** that simplify content creation.
- **Community-driven incentive mechanisms**, including voting, reputation, and collaborative moderation.
- **Cross-community integration** connecting Turkic-speaking populations within a shared digital ecosystem.

Together, these principles provide a practical framework for improving digital knowledge production and information accessibility. The following section presents the **Andisha** platform as a conceptual case study demonstrating how these principles can be applied in practice.

To illustrate the proposed framework, this study presents **Andisha** as a conceptual case study for addressing digital content scarcity in Turkic languages through a multilingual, community-driven platform.

Currently in the **Minimum Viable Product (MVP)** stage, Andisha is designed around four principles identified in this study: multilingual infrastructure, low barriers to participation, community-driven content creation, and cross-community collaboration.

The platform supports multiple writing systems through an internationalization (i18n) architecture and enables users to create and evaluate content using communities, posts,

comments, and voting mechanisms. Its primary objective is to encourage knowledge sharing and increase the amount of searchable digital content across Turkic languages. Unlike platforms focused on individual countries or languages, Andisha aims to connect Turkic-speaking communities within a shared digital environment. Future development includes AI-assisted translation, automated moderation, and recommendation systems to further improve information accessibility.

Overall, Andisha represents a **conceptual implementation** of the framework proposed in this study. Its effectiveness should be validated through future empirical research based on user participation and quantitative performance indicators.

Comparing of Andisha Network with other platforms

Comparison of Existing Community Platforms and the Proposed Andisha Platform

Feature	Wikipedia	Reddit	Stack Overflow	Andisha (Proposed Platform)
User-generated content	Yes	Yes	Yes	Proposed
Community discussions	Partial	Yes	Partial	Proposed
Voting mechanism	Partial	Yes	Yes	Proposed
Knowledge sharing	Yes	Partial	Yes	Proposed
Designed specifically for Turkic languages	No	No	No	Proposed
Primary objective: Increasing digital content in Turkic languages	No	No	No	Proposed
Support for multilingual Turkic communities	Partial	Partial	No	Proposed
Potential contribution to search engine indexing	Yes	Partial	Yes	Proposed

Source: Prepared by the author based on analysis of platform functionality.

Conclusion

This study examined the impact of digital content scarcity in Turkic languages on search engine performance through statistical evidence and recent research on low-resource languages. The findings indicate that although more than **200 million people** speak a Turkic language, these languages remain significantly underrepresented in the global digital ecosystem, limiting information accessibility and reducing search effectiveness. [1][3]

The study further demonstrates that improving search quality requires more than advances in search engine algorithms. Sustainable progress depends on increasing the availability of high-quality, user-generated digital content that can be effectively indexed and processed by modern information retrieval systems. Community-driven knowledge-sharing platforms provide a practical approach to achieving this objective.

As a conceptual case study, the **Andisha** platform illustrates how multilingual infrastructure, community participation, and cross-community collaboration may contribute to expanding digital knowledge resources for Turkic-speaking communities. Although the platform is still under development, it demonstrates how the theoretical principles discussed in this research can be translated into a practical digital environment.

Future research should evaluate this approach through empirical studies involving real users, quantitative measurements of content growth, search engine indexing performance, and AI-assisted multilingual technologies. Such investigations would provide stronger evidence regarding the long-term effectiveness of community-based platforms in improving digital knowledge accessibility for low-resource languages.

References:

1. **[1]** W3Techs, *Usage Statistics and Market Share of Content Languages for Websites*, 2026. Available: https://w3techs.com/technologies/overview/content_language
2. **[2]** S. Tukenov *et al.*, "SozKZ: Training Efficient Small Language Models for Kazakh from Scratch," *arXiv preprint arXiv:2603.20854*, 2026. Available: https://arxiv.org/pdf/2603.20854
3. **[3]** Wikipedia Contributors, "Turkic Languages," *Wikipedia, The Free Encyclopedia*. Available: https://en.wikipedia.org/wiki/Turkic_languages
4. **[4]** United States Patent and Trademark Office (USPTO), *Structured Cross-*

Lingual Relevance Feedback for Enhancing Search Results, Patent No. US8645289. Available: https://image-ppubs.uspto.gov/dirsearch-public/print/downloadPdf/8645289

5. **[5]** "Bridging the Gap: A Survey of Document Retrieval Techniques for High-Resource and Low-Resource Languages," *Journal of Network and Computer Applications*, Elsevier, 2025. Available: https://www.sciencedirect.com/science/article/abs/pii/S1574013725000322
6. **[6]** "Bridging Language Gaps: Advances in Cross-Lingual Information Retrieval with Multilingual Large Language Models," *arXiv preprint*, 2025.
7. **[7]** S. D. Pachpande and P. U. Bhalchandra, "Cross Language Information Retrieval Based on Automatic Query Translation," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, 2023.
8. **[8]** "What Are the Limits of Cross-Lingual Dense Passage Retrieval for Low-Resource Languages?" *arXiv preprint*, 2024.
9. **[9]** H. H. Nigatu, J. Canny, and S. E. Chasins, "Low-Resourced Languages and Online Knowledge Repositories: A Need-Finding Study," *arXiv preprint arXiv:2405.16669*, 2024. Available: https://arxiv.org/pdf/2405.16669
10. **[10]** Harvard Digital Innovation and Transformation, "Wikipedia, the Father of Crowdsourcing," 2015. Available: https://d3.harvard.edu/platform-digit/submission/wikipedia-the-father-of-crowdsourcing/
11. **[11]** Crowdsourcing Week, "Co-Creating Content Through Crowdsourcing Boosts Valuable Diversity," 2025. Available: https://crowdsourcingweek.com/blog/crowdsourcing-co-created-content/