Kashmiri Language in the Age of Artificial Intelligence

Gazi Imtiyaz Ahmad[1], Syed Ishfaq Manzoor[2]

Abstract

The Kashmiri language, spoken primarily in the Kashmir Valley, represents a rich cultural and linguistic heritage. One of the oldest languages still in use on the subcontinent, Kashmiri is the only writing branch of the Dardic group of Indo-Aryan languages. It is renowned for its adaptability and its intricate history of Persian, Sanskrit, and Central Asian influences. However, its presence in the rapidly evolving field of Artificial Intelligence remains limited. In the era of artificial intelligence, very little or no work has been done on the Kashmiri language. This paper explores the current state, challenges, and potential opportunities for integrating Kashmiri language processing within AI frameworks. It emphasizes the importance of developing natural language processing tools, speech recognition systems, and machine translation models tailored to Kashmiri, thereby contributing to language preservation and digital inclusion.

Keywords: Artificial Intelligence, Natural Language Processing, Machine Translation, Speech Recognition.

1. INTRODUCTION

Artificial Intelligence (AI) has revolutionized the processing and understanding of human languages, enabling unprecedented advancements in communication, information access, and technology-driven services across diverse domains. The rapid evolution of AI technologies, particularly in natural language processing (NLP), speech recognition, and machine translation, has transformed how languages are digitized, analysed, and utilized. While widely spoken global languages such as English, Mandarin, and Spanish benefit from extensive AI research, development, and deployment, many regional and minority languages remain marginalized in this technological transformation. This disparity results in a significant digital divide, where speakers of less-resourced languages face limited access to AI-powered tools and services that are increasingly integral to education, governance, healthcare, and social interaction. [1]

The Kashmiri language, spoken primarily in the Kashmir Valley and among diaspora communities, exemplifies such underrepresentation within the AI landscape. Kashmiri is a language with a rich cultural and historical heritage, reflecting centuries of literary tradition, oral storytelling, and unique linguistic features. Despite its significance, Kashmiri has seen limited integration within AI frameworks, which restricts the availability of advanced language technologies for its speakers. This underrepresentation poses significant risks to the vitality and sustainability of the language in the digital era, where the dominance of major languages threatens linguistic diversity and cultural identity. [2]

The digital divide affecting Kashmiri is multifaceted, encompassing several interrelated challenges. One major issue is the scarcity of digitized Kashmiri language resources, including annotated corpora, audio recordings, and parallel texts, which are essential for training AI models. Additionally, the lack of standardized orthography complicates text processing, as Kashmiri is written in multiple scripts—primarily Perso-Arabic and Devanagari—with variations in usage depending on region and community. This script variability introduces complexities in text normalization, tokenization, and model training. Furthermore, Kashmiri exhibits considerable dialectal diversity, with regional variations in

phonetics, vocabulary, and syntax, which challenge the development of generalized AI models capable of accurately representing the language's full spectrum.[3]

These linguistic and resource constraints hinder the development of essential AI applications such as NLP tools, speech recognition systems, and machine translation models tailored specifically to Kashmiri. The absence of such technologies limits educational, social, and technological opportunities for Kashmiri speakers, constraining their participation in the digital world and exacerbating existing inequalities. Moreover, the lack of AI integration impedes efforts toward the preservation and revitalization of Kashmiri, as digital tools increasingly shape language use, transmission, and visibility [4]. Addressing these challenges requires a focused, multidisciplinary effort to develop AI technologies that respect and incorporate Kashmiri's unique linguistic features. This includes accounting for its rich phonetic inventory, complex morphological structures, and script variability. Developing robust AI models for Kashmiri necessitates comprehensive corpus development, community engagement, and innovative methodological approaches such as transfer learning from related languages. Integrating Kashmiri into AI systems not only supports language preservation but also aligns with broader goals of linguistic diversity, cultural inclusion, and equitable access to technology.[5]

This paper aims to provide an in-depth exploration of the current status of Kashmiri language integration within AI, identifying key challenges and highlighting potential opportunities for advancing Kashmiri language technologies. Through a comprehensive examination of Kashmiri's linguistic characteristics, existing AI efforts, and methodological frameworks, the study underscores the critical need for collaborative, culturally informed AI research. Such efforts will empower Kashmiri speakers by expanding digital inclusion and contribute to enriching the global AI language landscape with greater linguistic diversity and representation.

2. LINGUISTIC FEATURES OF KASHMIRI

Kashmiri belongs to the Dardic subgroup of the Indo-Aryan languages and exhibits unique phonetic, morphological, and syntactic characteristics. Its rich vowel inventory, use of the Perso-Arabic script alongside Devanagari, and complex verb conjugations present specific challenges for computational modeling. Understanding these features is crucial for developing effective AI-based language technologies. About 7 million people mostly in Kashmir valley are native speakers of Kashmiri language. A sub-group of Indo-Aryan languages, Kashmiri languages is classified under the Dardic sub-group which makes it quite different from other North Indian languages like Hindi, Punjabi etc. however, owing to its geographical location and historical evolution, Kashmiri language has preserved archaic linguistic features [6]. The language exhibits several distinctive linguistic features:

Phonology: Kashmiri has a rich consonant inventory, including voiced, voiceless, aspirated, and un-aspirated stops. It features a set of retroflex consonants common in South Asian languages. Kashmiri is notable for its use of vowel harmony and a system of front rounded vowels, which is rare among Indo-Aryan languages. It has a tonal system, with high and low tones that can distinguish meaning between words.

Morphology: Kashmiri is a fusional language with inflectional morphology for nouns, pronouns, verbs, and adjectives. It employs gender (masculine and feminine), number (singular and plural), and case marking (nominative, accusative, genitive, dative, ablative, locative, and instrumental). The verb system encodes tense, aspect, mood, person, and number, with complex conjugation patterns. Postpositions are used instead of prepositions to indicate grammatical relations.

Syntax: The typical word order is Subject-Object-Verb (SOV). Kashmiri exhibits ergative alignment in the past tense, where the subject of a transitive verb is marked differently from the subject of an intransitive verb. It allows for relatively free word order due to its rich case marking system. It uses auxiliary verbs extensively to express tense and aspect.
Lexicon: Kashmiri vocabulary includes a significant number of loanwords from Sanskrit, Persian, Arabic, and more recently, Urdu and English. It retains many archaic Indo-Aryan lexical items not found in other modern Indo-Aryan languages.
Writing System: Kashmiri is written in multiple scripts: primarily the Perso-Arabic script (Nastaliq style), but also Devanagari and the Sharada script historically.The Perso-Arabic script has been adapted to represent Kashmiri phonology, including vowels and tonal distinctions.
Sociolinguistic Features: Kashmiri displays diglossia, with different registers used in formal and informal contexts. It is influenced by multilingualism in the region, often mixing with Urdu, Hindi, and English in daily use.

3. CURRENT STATUS OF KASHMIRI IN AI

Due to a lack of resources and datasets, the Kashmiri language—which is acknowledged as one of the low-resource languages—has a rich cultural legacy but is still unexplored in NLP.
The lack of digitized data, corpora, and tools makes Kashmiri language as a low-resource language in the context of AI/NLP. This is reflected in limited representation in AI systems.
The main aspects include:
Data Scarcity: Kashmiri has limited large-scale, high-quality digital corpora, which constrains the development of robust AI models for tasks like natural language processing (NLP), speech recognition, and machine translation.
Script Diversity: The use of multiple scripts (Perso-Arabic, Devanagari, Sharada) complicates text processing and standardization efforts in AI, requiring script-specific models or transliteration systems.
Phonological Complexity: The tonal system and unique vowel harmony in Kashmiri present challenges for speech synthesis and recognition technologies, necessitating specialized acoustic models.
Morphosyntactic Features: The fusional morphology, ergative alignment, and flexible word order demand sophisticated syntactic parsers and morphological analyzers to accurately model Kashmiri grammar in AI applications.
Multilingual and Code-Switching Context: Frequent mixing with Urdu, Hindi, and English in daily use requires AI systems to handle code-switching and mixed-language input effectively.
The application of AI to Kashmiri language processing is nascent. Limited digital corpora, scarcity of annotated datasets, and lack of standardized orthography hinder the development of robust AI models. Existing efforts mainly include small-scale projects on text digitization and rudimentary machine translation prototypes, often relying on transfer learning from related languages. However, the modern datasets, benchmark studies, and Kashmiri-centric AI applications have been developed in recent years. Though still far behind major Indic languages, these efforts represent a change from near-absence in computational linguistics to an emerging ecosystem of tools and resources. Lone et. al 2022 in [7] lists the available Kashmiri-English dictionaries, Kashmiri Wordnet, EMILLE monolingual corpora, a spell checker, and some speech tools, among other NLP resources, and identified significant research gaps. The first systematic baseline for Kashmiri text classification was reported in [8] with the release of a 15,036-snippet multiclass news dataset that was used to assess machine learning, deep learning, transformers, and LLMs. Neural machine translation (MT)

is possible but limited by parallel data size and domain coverage, according to recent research on Kashmiri–English machine translation using deep neural architectures by [9.] Researches such as reported by [10-11] and [12] focused on morphology, sense-annotated corpora, and word-sense disambiguation, which serve as building blocks for more complex downstream systems. An independent engineer has created a Kashmiri-focused language assistant known as "KashmiriGPT," which is specifically designed as a language-preservation tool that enables users to communicate with an AI system in Kashmiri, Roman Kashmiri, and English. More extensive attempts to teach AI models Indian dialects, including Kashmiri, by gathering local data and creating domain-specific models are detailed in media publications and startup projects.

4. CHALLENGES IN KASHMIRI AI INTEGRATION

Kashmiri's integration into AI faces multifaceted challenges stemming from its low-resource status, unique linguistic properties, and structural barriers. These obstacles hinder the development of reliable NLP tools, machine translation, speech systems, and other applications essential for digital inclusion [13].

Data Scarcity and Quality: Limited availability of large, high-quality datasets in the Kashmiri language hinders effective training of AI models. The lack of digitized and annotated text corpora, speech data, and domain-specific resources restricts the development of robust natural language processing (NLP) systems [14]. Partner with local academic institutions, language experts, and communities to gather diverse Kashmiri text and speech data.

Linguistic Complexity: Kashmiri's rich morphology, complex syntax, and use of multiple scripts (Perso-Arabic and Devanagari) pose significant challenges for AI algorithms. Handling dialectal variations and context-dependent meanings requires sophisticated linguistic modelling [15]. Partner with local academic institutions, language experts, and communities to gather diverse Kashmiri text and speech data.

Script and Orthography Variability: The coexistence of multiple writing systems complicates text normalization, tokenization, and model training. AI systems must accommodate script conversion or script-specific models, increasing development complexity. Multiple scripts complicate text normalization and processing.

Limited Technological Infrastructure: The region's limited access to advanced computing resources can affect data collection, model training, and deployment of AI applications tailored for Kashmiri [16]. Utilization of cloud-based platforms to mitigate local hardware limitations and enable scalable model training can overcome this limitation.

Cultural and Contextual Nuances: AI models often struggle to incorporate cultural references, idiomatic expressions, and local context unique to Kashmiri speakers, reducing the relevance and accuracy of generated content or responses. Diverse dialects within Kashmiri affect model generalizability

Lack of Domain-Specific Tools and Benchmarks*:* Absence of standardized evaluation benchmarks and domain-specific AI tools for Kashmiri limits the ability to measure progress and compare model performance effectively [17]. Develop crowdsourcing platforms or community-driven initiatives for data annotation, focusing on dialectal and script diversity. Create comprehensive lexical databases, morphological analyzers, and syntactic parsers tailored to Kashmiri.

Awareness and Adoption Barriers: Limited awareness among researchers, developers, and the community about AI possibilities in Kashmiri restricts collaboration and resource sharing, slowing down innovation and integration efforts. Develop script-aware and dialect-sensitive AI models that can handle multiple orthographies and linguistic complexities. Incorporate

cultural and contextual knowledge by embedding local idioms, references, and socio-cultural norms into model training data.

To advance Kashmiri AI integration, addressing these challenges through collaborative data collection, development of linguistic resources, infrastructure improvement, and culturally informed model design is essential.

5. CONCLUSION

Integrating AI for the Kashmiri language presents a unique set of challenges rooted in data scarcity, linguistic complexity, script variability, and infrastructural limitations. Addressing these issues requires a multifaceted strategy encompassing collaborative data collection, development of specialized linguistic resources, and deployment of script- and dialect-aware AI models. Enhancing technological infrastructure and fostering community engagement are equally critical to ensure sustainable progress. Ethical considerations, including data privacy and socio-political sensitivities, must guide the development and deployment of AI systems to maintain cultural relevance and trust. Through coordinated efforts focusing on these strategic areas, AI integration can significantly advance the preservation, accessibility, and technological empowerment of the Kashmiri language, ultimately contributing to its digital revitalization and broader inclusion in the global AI landscape.

REFERENCES

1. Lawaye, A. A., Mir, T. A., Mir, M. H., & Ahmed, G. (2024). Machine Learning Approach for Kashmiri Word Sense Disambiguation. In *Empowering Low-Resource Languages With NLP Solutions* (pp. 113-136). IGI Global Scientific Publishing
2. Lone, N. A., Giri, K. J., & Bashir, R. (2023). Machine intelligence for language translation from Kashmiri to English. *Journal of Information & Knowledge Management*, *22*(04), 2250074.
3. Qumar, S. M. U., Azim, M., & Quadri, S. M. K. (2025). Emerging resources, enduring challenges: a comprehensive study of Kashmiri parallel corpus. *AI & SOCIETY*, *40*(4), 2385-2403.
4. Deyar, D. U., Ramani, A., Gupta, D., Nair, P. C., & Venugopalan, M. (2025). Dataset creation and benchmarking for Kashmiri news snippet classification using fine-tuned transformer and LLM models in a low resource setting. *Scientific Reports*, *15*(1), 40828.
5. Mir, T. A., & Lawaye, A. A. (2025). Word sense disambiguation corpus for Kashmiri. *Natural Language Processing*, *31*(2), 631-654
6. Asher, R. E. (1970). A Reference Grammer of Kashmiri. By Braj B. Kachru. Urbana: Department of Linguistics, University of Illinois, 1969. xxv, 416 pp. Appendix, Glossary, Indexes, np. *The Journal of Asian Studies*, *29*(4), 977-977
7. Lone, N. A., Giri, K. J., & Bashir, R. (2022). Natural Language Processing Resources for the Kashmiri Language. *Indian Journal of Science and Technology*, *15*(43), 2275-2281
8. Malik, H. N. (2026). ks-lit-3m: A 3.1 million word kashmiri text dataset for large language model pretraining. *arXiv preprint arXiv:2601.01091*.
9. Qumar, S. M. U., Azim, M., & Quadri, S. M. K. (2024). Addressing the data gap: building a parallel corpus for Kashmiri language. *International Journal of Information Technology*, *16*(7), 4363-4379
10. Benish Afzal Want Computational Modeling of Kashmiri Morphology: Finite-State Analysis and Linguistic Insights International Journal of Humanities Social Science and Management (IJHSSM) Volume 4, Issue 2, Mar.-Apr., 2024, pp: 589-601

11. Mir, T. A., & Lawaye, A. A. (2025). Word sense disambiguation corpus for Kashmiri. *Natural Language Processing*, *31*(2), 631-654
12. Mir, T. A., Lawaye, A. A., Rana, P., & Ahmed, G. (2023). Building kashmiri sense annotated corpus and its Usage in supervised word sense disambiguation. *Indian Journal of Science and Technology*, *16*(13), 1021-1029
13. Sikandar, M. (2025). The Morphology of Kashmiri: Structure, Formation, and Linguistic Influence. *Journal of Political Stability Archive*, *3*(4), 1130-1140.
14. Ahmad, S., Jamshaid, K., & Ammar, A. (2023). A Corpus-based Study of Newspaper Articles on Lockdown Issue in Indian Occupied Kashmir. *Corporum: Journal of Corpus Linguistics*, *6*(1), 22-37.
15. Bhat, M. A. (2017). *The changing language roles and linguistic identities of the Kashmiri speech community*. Cambridge Scholars Publishing.
16. Malik, H. N. KashScript: Revitalizing Kashmiri through a Comprehensive Transliteration Framework for Rescuing Kashmiri from Digital Extinction-Integrating InPage-to-Unicode Conversion, ALA-LC Romanization, and AI-Ready.
17. Giri, K. J., Lone, N. A., Bashir, R., & Bhat, J. I. (2024, February). English Kashmiri Machine Translation System related to Tourism Domain. In *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1713-1717). IEEE.