International Journal of Multidisciplinary Research and Growth Evaluation.

Machine-Readable Malayalam Dictionaries and Lexical Databases

Santhosh M
Assistant Professor, School of Linguistics, Thunchath Ezhuthachan Malayalam University, Tirur, Malappuram, Kerala, 676502, India

* Corresponding Author: **Santhosh M**

Article Info

ISSN (Online): 2582-7138
Impact Factor (RSIF): 8.04
Volume: 01
Issue: 05
Received: 04-11-2020
Accepted: 06-12-2020
Page No: 1177-1182

Abstract
The rapid growth of Natural Language Processing (NLP) and language technologies has increased the importance of structured lexical resources for Indian languages. Malayalam, a morphologically rich Dravidian language spoken predominantly in Kerala, presents unique challenges and opportunities in computational linguistics. Among the foundational resources required for language technology development are Machine-Readable Dictionaries (MRDs) and Lexical Databases. This article examines the development, structure, significance, and applications of machine-readable Malayalam dictionaries and lexical databases. It discusses the evolution of Malayalam computational lexicography, major projects and initiatives, linguistic challenges, and the role of lexical resources in language processing systems. The study also highlights the relationship between traditional lexicography and computational approaches and demonstrates how digital tools build upon centuries of traditional Malayalam scholarship.

DOI: https://doi.org/10.54660/.IJMRGE.2020.1.5.1177-1182

Keywords: Computational Lexicography, Language Technology, Lexical Databases, Machine-Readable Dictionaries (MRDs), Natural Language Processing (NLP)

1. Introduction

Lexicography has traditionally focused on the compilation and organization of dictionaries for human use. With the advent of computational linguistics, dictionaries began to be digitized and structured for machine processing, leading to the emergence of Machine-Readable Dictionaries (MRDs). Lexical databases evolved as sophisticated lexical resources capable of representing semantic and grammatical relationships among lexical items (Fellbaum, 1998). Malayalam, one of the major Dravidian languages of India, possesses a rich literary tradition and a complex morphological structure. Computational treatment of Malayalam requires organized lexical resources capable of supporting applications such as machine translation, information retrieval, spell checking, speech processing and text analysis. Considerable efforts have been made in developing Malayalam lexical resources, including electronic dictionaries, morphological analyzers, WordNet projects, annotated corpora, and open-source lexical databases. However, compared with resource-rich languages such as English, Malayalam computational lexicography remained in a developmental stage. This article analyzes the nature and development of machine-readable Malayalam dictionaries and lexical databases within the broader framework of computational lexicography.

2. Literature Review

The development of Machine-Readable Dictionaries (MRDs) and lexical databases has emerged as an important area within computational linguistics and Natural Language Processing (NLP). Early theoretical foundations for electronic lexical resources were established through works such as Fellbaum's WordNet: An Electronic Lexical Database [6], investigated how lexical items could be represented through semantic networks rather than simple alphabetical listings. The formal relationships between words (different types of connections among words) can be used to convert a printed dictionary into a structured digital database [16].

In the Indian context, the Government of India's Technology Development for Indian Languages (TDIL) programme played a significant role in encouraging lexical resource development for Indian languages. Research initiatives under TDIL, C-DAC, and Indian language NLP projects highlighted the necessity of machine-readable lexicons, annotated corpora, and morphological analyzers for computational processing of morphologically rich languages. Malayalam computational lexicography developed gradually through both institutional and community-driven initiatives. Research publications on Malayalam morphology and NLP repeatedly pointed out the challenges posed by Malayalam's agglutinative structure, sandhi processes, and dialectal variation. These studies emphasized the need for structured lexical databases capable of supporting morphological analysis, POS tagging, machine translation, and information retrieval.

Several important Malayalam lexical resources emerged such as Malayalam WordNet, developed under the IndoWordNet framework, introduced semantic lexical networking for Malayalam through synsets and semantic relations such as synonymy, hypernymy, and antonymy. Simultaneously, digitization projects involving Dr. Hermann Gundert's A Malayalam and English Dictionary [9], Sabdataravali and the Olam lexical database demonstrated the transition from traditional print lexicography to searchable machine-readable lexical systems.

Open-source initiatives such as Swathanthra Malayalam Computing (SMC) further expanded Malayalam computational lexicography through morphology analyzers, spell-checkers, lexical databases, and Unicode-based language computing tools [2]. Malayalam Wiktionary and collaborative lexical platforms also contributed to the growth of publicly accessible lexical resources useful for NLP experimentation and digital lexicography.

Although considerable work had been carried out in Malayalam language technology, studies specifically examining Malayalam MRDs and lexical databases in a unified computational lexicographic framework remained limited. Most existing studies focused separately on morphology, machine translation, WordNet development, or digital dictionaries. Therefore, the present article attempts to provide a consolidated study of Malayalam Machine-Readable Dictionaries and lexical databases by examining their development, structure, linguistic features, and computational significance within the broader field of Malayalam computational lexicography.

3. Lexical Resources and Computational Lexicography

3.1. Lexical Resources

Lexical resources are structured repositories containing linguistic information about words and lexical units. These include dictionaries, thesauri, corpora, ontologies, lexical databases and WordNets among others. Lexical resources serve as essential infrastructure for both linguistic research and NLP applications.

3.2. Computational Lexicography

According to Jackson, computational lexicography involves the application of computational methods to the creation, management, and utilization of lexical resources [7]. It involves digitizing dictionaries, building lexical databases, and utilizing automated corpus analysis for morphological, semantic, and structural processing. In the Malayalam context, computational lexicography emerged primarily through initiatives in language digitization, language technology projects, academic NLP research, and open-source Malayalam computing movements.

4. Machine-Readable Dictionaries (MRDs)

4.1. The Digital Lexicon

A Machine-Readable Dictionary (MRD) is a dictionary stored in electronic format and structured in a way that allows computational access and processing. Unlike physical books, MRDs support digital tasks like automated queries, grammatical breakdowns, and computer-based data extraction.

4.2. Characteristics of MRDs

Machine-readable dictionaries generally contain: Headwords, pronunciation, grammatical category, definitions, usage examples, semantic labels and inflectional information. The electronic representation may use: XML, SGML, Unicode encoding, database structures, or JSON and CSV formats. MRDs are not merely digital versions of printed dictionaries; rather, they function as structured linguistic resources that can be processed computationally for various NLP and language technology applications.

5. Development of Malayalam Machine-Readable Dictionaries

5.1. Early Digitization Efforts

Initial Malayalam dictionary digitization efforts were largely motivated by language preservation, educational software, spell-checking tools and Malayalam computing initiatives [17]. The adoption of Unicode for Malayalam script significantly accelerated computational processing and dictionary digitization. One of the earliest pioneers in Malayalam lexical digitization was Datuk K. J. Joseph, who digitized Malayalam dictionaries and literary materials during the late 1990s. His digitized lexical corpus later became foundational for several Malayalam digital dictionary projects.

5.2. Institutional Contributions

Several institutions and organizations are making significant contributions to the development of digital Malayalam lexical resources.

Centre for Development of Advanced Computing (C-DAC) played a major role in Indian language computing through font standardization, text processing tools, electronic lexical resources, morphological analyzers, spell-checking systems and language technology development [12-14]. Universities and research centers contributed through computational linguistics research, corpus creation, morphological analysis, bilingual dictionary development and semantic resource construction. Institutions such as IIIT Hyderabad and Amrita Vishwa Vidyapeetham contribute to Indian language NLP projects, including lexical and morphological resource development relevant to Malayalam.

5.3. TDIL (Technology Development for Indian Languages)

National projects under the TDIL programme supported: lexical databases, machine translation systems, morphological analyzers, annotated corpora, bilingual lexical resources and WordNet development. These initiatives have

provided and continue to provide foundational infrastructure for Malayalam NLP and computational lexicography. (11)

5.4. Community-driven Open-source Initiatives

Apart from government-supported and academic initiatives, community-driven projects also played an important role in Malayalam computational lexicography. One of the most influential initiatives was SMC, which emerged from the Malayalam free-software movement during the early 2000s. SMC contributed significantly to: Malayalam dictionaries and lexical resources, morphology analyzers, spell-checking tools, computational lexicons, Unicode Malayalam computing and the SILPA language processing framework [10]. The morphology analyzer projects associated with SMC included: root-word databases, grammatical category information, inflectional analysis systems and lexical parsing support.These resources became important for Malayalam NLP experimentation and open-source language technology development.

5.5. Datuk Corpus and Olam Lexical Database

The lexical database behind Olam originated primarily from the digitization efforts of Datuk K. J. Joseph. Later, the digitized lexical corpus was integrated into the Olam project by Kailash Nadh [1]. Olam evolved into a fully searchable, open-source English–Malayalam dictionary platform providing machine-readable lexical data for both public use and computational processing. Unlike static web dictionaries, Olam provides structured lexical datasets in formats like JSON, CSV and database dumps. These structured resources enabled:

NLP model training, lexical parsing, translation APIs, spell-checking systems and third-party dictionary applications. Olam thus became one of the most important publicly accessible machine-readable lexical resources for Malayalam.

5.6. Digitized Traditional Dictionaries

The transition from traditional Malayalam dictionaries to digital machine-readable formats marked an important stage in Malayalam computational lexicography. Major Malayalam dictionaries such as Dr. Hermann Gundert's A Malayalam and English Dictionary and

Sabdatharavali were digitized and transformed into searchable lexical databases.

5.6.1. Dr. Hermann Gundert's A Malayalam and English Dictionary

The digitized and searchable versions of Gundert's dictionary became available through archival and academic digitization projects such as the Digital Dictionaries of South Asia (DDSA). These digital versions allowed: exact word searches, substring searches, lexical retrieval and computational parsing. The backend structure employed OCR processing and metadata tagging, enabling machine-readable access to lexical information including definitions, lexical roots and grammatical categories. Thus, Gundert's dictionary evolved from a traditional print dictionary into a computationally accessible lexical resource.

5.6.2. Sabdatharavali

Sabdatharavali was digitized and made available as a searchable online lexical platform through initiatives such as the Sayahna Foundation. The digitized version employed Unicode Malayalam encoding and searchable database structures. Lexical hosting through platforms such as Lexonomy further increased its machine-readability and computational accessibility.

These digitization projects transformed traditional Malayalam lexicography into machine-processable lexical infrastructure.

6. Structure of Malayalam MRDs

Malayalam MRDs generally include several layers of linguistic information.

6.1. Orthographic Information

Malayalam script representation, spelling variants, transliteration and Unicode normalization.

6.2. Morphological Information

Malayalam is highly agglutinative. Therefore, MRDs often include root forms, suffix information, inflectional paradigms and derivational patterns.

6.3. Syntactic Information

Part-of-speech labels, grammatical features and subcategorization information.

6.4. Semantic Information

Definitions, synonymy, antonymy, hyponymy, meronymy, and semantic domains.

6.5. Bilingual and Multilingual Information

Many Malayalam MRDs are bilingual or multilingual such as Malayalam - English,

English - Malayalam and English - English - Malayalam - Hindi. These support translation systems and multilingual NLP applications.

7. Lexical Databases

7.1. Core Concept

A lexical database is a structured repository of lexical information organized for computational access and semantic representation. Unlike ordinary electronic dictionaries, lexical databases represent semantic relations, conceptual hierarchies, lexical networks and ontological structures.

8.1 Difference Between MRDs and Lexical Databases

Table 1

Feature	MRDs	Lexical Databases
Primary function	Electronic dictionary	Structured lexical knowledge
Organization	Entry-based	Relation/ network-based
Semantic relations	Limited	Extensive
NLP suitability	Moderate	High
Knowledge representation	Basic	Advanced

9. Malayalam Lexical Databases

9.1. WordNet

One of the most important Malayalam lexical database initiatives was the Malayalam WordNet, developed as part of the larger IndoWordNet and Dravidian WordNet projects. Inspired by Princeton WordNet, Malayalam WordNet organized lexical items into synsets,

semantic categories and lexical / relation networks. These structures are built upon major

semantic relations such as synonymy, hypernymy, hyponymy, meronymy and antonymy.

Fellbaum (1998) explains that WordNet organizes lexical items into semantic synsets and relational networks, a model later adopted inSand IndoWordNet projects [16].

Malayalam provided XML and JSON interfaces for computational processing. Malayalam WordNet became important for: machine translation, semantic search, word sense disambiguation, question answering systems, semantic analysis and NLP applications.

9.2. IndoWordNet

IndoWordNet connected Malayalam lexical data with lexical resources from other Indian languages [5]. Each lexical entry typically included: synonym sets, glosses and example sentences. IndoWordNet became important for cross-lingual NLP, bilingual dictionary development, semantic mapping and machine translation.

10. Malayalam Wiktionary and Collaborative Lexical Resources

Malayalam Wiktionary emerged as a collaborative machine-readable lexical resource maintained by the Malayalam Wikimedia community. It provided meanings, etymology, grammatical information and pronunciation. Though less formally structured than WordNet, Malayalam Wiktionary became useful as: an open lexical database, corpus material, a collaborative lexicographic platform and an experimental resource for NLP research.

11. Morphological Lexicons and Computational Resources

Finite-state transducer-based approaches have demonstrated that Malayalam morphological analyzers depend heavily on structured lexical databases containing root forms, inflectional paradigms, and grammatical information [3]. To support such computational processing, Malayalam morphological lexicons and analyzers were developed by institutions such as C-DAC, IIIT Hyderabad, and participating universities under the TDIL programme. These resources provided essential linguistic information for morphological analysis, part-of-speech (POS) tagging, parsing, spell checking, and machine translation. By incorporating root-word databases, grammatical categories, and inflectional and derivational patterns, these computational lexicons became foundational resources for Malayalam computational morphology and Natural Language Processing (NLP) systems.12. Linguistic Challenges in Malayalam Lexical Resource Development

11.1. Morphological Complexity

Malayalam is considered morphologically complex because words are formed by attaching several grammatical elements to a root word. This characteristic mainly results from its Agglutinative nature, Inflectional system, Derivational processes, Compound word formation

These features allow a single lexical root to generate a very large number of word forms.

Agglutinative nature words: കുട്ടികൾ, കുട്ടികളെ, കുട്ടികളോടൊപ്പം.

Inflectional system words: പോകുന്നു, പോകും, പോകുമ്പോൾ.

Derivational processes words: പഠനം, പഠിത്താവ്, പഠിക്കൽ.

Compound word formation words: പൂന്തോട്ടം, താമരക്കണ്ണൻ നീലത്താമര, ആനക്കൊമ്പ്.

Due to Malayalam's agglutinative morphology, a single lexical root can generate numerous surface forms through suffixation and inflection [4]. This complexity complicates dictionary structuring and lexical parsing.

11.2. Sandhi and Compound Formation

Phonological changes during word combination create challenges for tokenization, lexical lookup, stemming and morphological analysis. As noted by Vinod *et al*. (2012), the complex morphophonemic rules of Malayalam create substantial hurdles for automated lexical lookup and morphological analysis [11].

11.3. Dialectal Variations

Malayalam exhibits considerable regional and social dialectal variation [18]. Many lexical items used in local dialects differ significantly from their standard Malayalam equivalents. This variation poses a challenge for the development of machine-readable dictionaries, lexical databases, and Natural Language Processing (NLP) applications.

Most Machine Translation (MT) systems, morphological analyzers, spell checkers, and lexical databases are primarily built on standard Malayalam vocabulary, orthography, and grammatical patterns. Consequently, when users enter dialectal forms, the system may fail to recognize the intended meaning because such forms are absent from its lexical resources.

For example, the standard Malayalam expression അച്ഛന്റെ വാഴത്തോട്ടം ('father's banana plantation') may occur in certain dialects as ബാപ്പച്ചീടെ ബായതോട്ടം. Similarly, standard forms such as നിങ്ങളുടെ കുട്ടികൾ എവിടെയാണ്? may appear in dialectal varieties as ങ്ങളെ കുട്ട്യള് ഏവ്ടെ? or ങ്ങളെ കുട്ട്യള് ഏടെ? Although these expressions convey essentially the same meaning, a system trained only on standard Malayalam may not identify the correspondence between lexical items such as വാഴ and ബായ.

Such dialectal variation complicates lexical lookup, morphological analysis, machine translation, and information retrieval, highlighting the need for broader dialectal coverage in Malayalam lexical resources.

11.4. Lack of Large Annotated Corpora

Malayalam suffers from limited tagged corpora, semantic annotations and parallel corpora. This can restrict advanced NLP development.

12. Applications of Malayalam MRDs and Lexical Databases

12.1. Natural Language Processing (NLP)

Malayalam MRDs and lexical databases became foundational resources for Morphological analysis, POS tagging, Parsing, Semantic analysis, Named Entity Recognition and Question - Answering systems

12.2. Machine Translation

Lexical databases support lexical mapping, semantic equivalence, bilingual alignment and cross-lingual processing.

12.3. Spell and Grammar Checking
Morphological and lexical data improve the error detection, correction suggestions and grammar analysis.

12.4 Speech Processing
Word databases make it easier for computers to generate human speech, understand spoken words and teach correct pronunciation.

12.5 Educational Technology and Digital Lexicography
Digital dictionaries and lexical systems support language learning, e-learning systems,
vocabulary development and digital lexicography. Thus, Malayalam MRDs served not only as lookup tools for users but also as foundational linguistic infrastructure for multiple computational and educational applications.

13. Relationship Between Traditional Lexicography and Computational Lexicography
Traditional Malayalam lexicography focused primarily on literary vocabulary, semantic explanation, etymology and language preservation. Computational lexicography introduced formal lexical structures, machine processability, semantic encoding, corpus-based analysis and computational querying. Machine-readable dictionaries often originated from printed dictionaries but required restructuring for computational purposes. Thus, computational lexicography did not replace traditional lexicography; rather, it extended and transformed it into machine-processable lexical systems.

14. Limitations of Malayalam Lexical Resources
Compared to English and some other Indian languages, Malayalam lexical resources remain relatively underdeveloped.

15. Future Prospects
A promising direction for future Malayalam language technology is the development of dialect-aware lexical resources and machine translation systems. Future machine-readable dictionaries and lexical databases may incorporate mappings between dialectal forms and their standard Malayalam equivalents. In addition, automatic dialect identification and dialect normalization modules could be integrated into NLP systems.
For instance, a dialect normalization layer could automatically convert a dialectal expression such as ബാപ്പച്ചീടെ ബായ തോട്ടം into its standard Malayalam equivalent അച്ഛന്റെ വാഴത്തോട്ടം before translation. The normalized form could then be translated accurately into English as father's banana plantation. Such developments would improve the ability of Malayalam NLP systems to process naturally occurring language beyond the standard variety and would make language technologies more inclusive of regional speech varieties.

16. Conclusion
Machine-readable Malayalam dictionaries and lexical databases constitute fundamental resources in Malayalam computational linguistics. While MRDs provide electronically accessible lexical information, lexical databases offer structured semantic and relational representations essential for advanced NLP applications. Malayalam computational lexicography has witnessed important developments through institutional initiatives, WordNet projects, digitized traditional dictionaries, open-source Malayalam computing movements and digital language technologies. Projects such as Olam, Malayalam WordNet,
IndoWordNet, SMC, digitized Gundert Dictionary, Sabdatharavali digitization and Malayalam Wiktionary demonstrated the gradual transition from traditional print lexicography to machine-processable lexical infrastructures [8].
Nevertheless, the field continued to face challenges arising from Malayalam's morphological richness, dialectal diversity, and limited annotated corpora. The transition from traditional lexicography to computational lexical systems marked a significant stage in the modernization of Malayalam language studies. The continued development of lexical resources remains essential for the advancement of Malayalam language technology, NLP research, AI-based language systems, and digital humanities research.

References

1. Nadh K. Alar: the making of an open-source dictionary. ZERODHA.tech. 2020 Sep 22.
2. Swathanthra Malayalam Computing. Research [Internet]. [cited 2026 Aug 1[7]. Available from: https://smc.org.in/en/research/
3. Thottingal S. Finite state transducer based morphology analysis for Malayalam language. In: Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages. European Association for Machine Translation; 2019. p. 1-5. Available from: https://aclanthology.org/W19-6801/
4. Mohanan KP. The theory of lexical phonology. Dordrecht: Reidel Publishing Company; 1986. Available from: http://tscheer.free.fr/interface/Mohanan%2086%20%20The%20Theory%20of%20Lexical%20Phonology
5. Rajendran S, Soman KP. Malayalam WordNet. Language in India. 2014;14(3):310-324. Available from: http://www.languageinindia.com/march2014/rajendran malayalamwordnetfinal
6. Soergel D. WordNet: an electronic lexical database. In: Fellbaum C, editor. WordNet: an electronic lexical database. Cambridge (MA): MIT Press; 1998. p. 131-161.
7. Jackson H. Lexicography: an introduction. London: Routledge; 2002.
8. Gundert H. Digitized versions of Gundert's textual materials [Internet]. University of Tübingen: OpenDigi. [cited 2026 Aug 1[7]. Available from: https://opendigi.ub.uni-tuebingen.de/opendigi/CiXIV68?ui_lang=eng
9. Gundert H. Aksharam. In: A Malayalam and English dictionary [Internet]. Digital Dictionaries of South Asia; University of Chicago; 1872. Available from: https://dsal.uchicago.edu/cgi-bin/philologic/getobject.pl?c.0:1:101.gundert
10. Thottingal S. A Malayalam WordNet [Internet]. SMC Discussion List; 2013 May 31. Available from: http://lists.smc.org.in/pipermail/discuss-smc.org.in/2013-May/015592.html
11. Vinod PM, Bhadran VK, Jayan V. Implementation of Malayalam morphological analyzer based on hybrid approach. Int J Comput Appl. 2012;46(12):24-28.

Available from: https://www.researchgate.net/publication/328066535_Implementation_of_Malayalam_Morphological_Analyzer_based_on_Hybrid_Approach
12. Centre for Development of Advanced Computing. Annual reports [Internet]. [cited 2026 Aug 1[7]. Available from: https://www.cdac.in/index.aspx?id=AnnualReports
13. Centre for Development of Advanced Computing. Multilingual computing [Internet]. [cited 2026 Aug 1[7]. Available from: https://www.cdac.in/index.aspx?id=mlingual
14. Centre for Development of Advanced Computing. Indian language fonts, corpora, dictionaries and tools [Internet]. [cited 2026 Aug 1[7]. Available from: https://cdac.in/index.aspx?id=printpage&print=mc_ilfcdt
15. Dancette J, L'Homme MC. Building specialized dictionaries using lexical functions. Linguistica Antverpiensia, New Series: Themes in Translation Studies. 2021;3:113-132. doi:10.52034/lanstts.v3i.107.
16. Fellbaum C, editor. WordNet: an electronic lexical database. Cambridge (MA): MIT Press; 1998.
17. Saidalavi C. Computer Bhashasastram. Tirur: Thunchath Ezhuthachan Malayalam University; 2020.
18. Sreenathan M. Bhashabhedapadahanam. Tirur: Thunchath Ezhuthachan Malayalam University; 2015.