==========
Language in India www.languageinindia.com ISSN 1930-2940 Vol. 25:11 November 2025
==========

Semantic Idiosyncrasies and Classification of Selected Multi-Word Expressions in Namboothiri Dialect

Mythili Raja
Research Scholar,
Department of Linguistics,
University of Kerala
mythili@keralauniversity.ac.in

Dr. Syam S.K.
Assistant Professor,
Department of Linguistics,
University of Kerala
drsyam@keralauniversity.ac.in

==========

Abstract

Multi-word expressions (henceforth MWEs) are an interesting feature of languages. They are expressions containing more than one word with a meaning, different from the meaning of their constituent elements. The meaning carried by these MWEs is specific. They are usually mastered and used extensively by native speakers. However, there are a few hurdles in understanding or decoding for non-native speakers. It becomes a bit more difficult when it comes to dealing with MWEs in a dialect of a language. The current study focuses on delving into the semantic idiosyncrasies and classification of MWEs in the Namboothiri dialect of Malayalam. This dialect is a caste dialect that carries its idiosyncrasies. Namboothiris are a subcategory of the Brahmins of Kerala. They are believed to have settled in the state as a result of the Aryan migration. They have influenced the development of Malayalam with their Sanskrit tradition that led to the maṇipRavāḷa tradition. The study is the first of its kind, as there are few or no studies about MWEs of a dialect of a language. MWE-based studies are very less even in the Malayalam language. Hence, the current study adds significantly to the existing knowledge of the MWEs of the Malayalam language.

Keywords: Multi-word expressions (MWEs), Malayalam, dialect, semantic idiosyncrasies, classification

Introduction

Multiword Expressions (MWE) can be described as "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag, Ivan A. et.al, 2002). They are lexical items that exhibit semantic idiosyncrasies. They can be classified into various types. Identification of these MWEs in a language is carried out based on certain properties and features. These properties are the basis of the semantic idiosyncrasies of MWEs. The research paper provides comprehensive information on the properties, types, and classification of MWEs.

The paper specifically focuses on analysing the semantic idiosyncrasies of MWEs in one of the caste dialects of the Malayalam language. The dialect is known as the Namboothiri dialect. Namboothiris, also known as Kerala brahmins or Malayala brahmins, are a community following the ancient Vedic tradition. According to Kesavan Veluthattu brahmin settlements of Kerala began around the end of the Sangam period in history. They brought in with them their tradition which is heavily influenced by Vedic tradition. The contribution of this community to the development of the Malayalam language is non-negligible because they influenced Malayalam language development with the Sanskrit tradition. The life and culture of this community is deeply entangled with the Sanskrit language. Hence, one can find a lot of idiosyncrasies in their dialect. Therefore, MWEs of this caste dialect is interesting with their uniqueness. Thus, in this paper the semantic idiosyncrasies and the classification of these MWEs are attempted.

Literature Review

The literature review carried out in relation to this study revealed that studies on Multi-Word Expressions (MWEs) in Malayalam are a less explored area. There are hardly any studies focusing on the MWEs in the dialects of Malayalam. The research paper by Treesa Anjaly Cyriac, Sobha Lalitha Devi (2022) is a work focusing on MWEs in Malayalam. In their paper, they are trying to explore multiwords in Malayalam and to classify them as per the three idiosyncrasies: semantic idiosyncrasy, syntactic idiosyncrasy, and statistic idiosyncrasy. The classification and features are given and are studied using Malayalam MWEs. Through this study, they identified how the linguistic features of Malayalam, such as agglutination, influenced its multiword expressions in terms of pronunciation and spelling. Malayalam has a set of code-mixed multiword expressions, which is also addressed in this study. Another

research work carried out in this area is the combined research paper by Jayan Vasudevan, V.K. Bhadran, and Arathi Raghunathan, 2016. This paper strives to establish the relevance of exploring the domain of Multi-Word Expressions in English and Malayalam. Next, the study carried out by Dhirendra Singh, Sudha Bhingardive, Pushpak Bhattacharyya creates MWEs annotation in Indian languages viz., Hindi and Marathi. Annotation is done for two types of MWEs: compound nouns and light verb constructions. The result of this study was that it created a resource which is made available publicly and can be used as a gold standard for Hindi and Marathi MWEs systems. Last but not the least, Mathieu Constant ATILF, Université de Lorraine & CNRS Gül¸ sen Eryi˘ git Istanbul Technical University Johanna Monti "L'Orientale" University of Naples Lonneke van der Plas University of Malta Carlos Ramisch§ Aix Marseille University, CNRS Michael Rosner University of Malta Amalia Todirascu LiLPa, Strasbourg University (2017) together worked with the aim not only to provide a focused review of MWE processing, but also to clarify the nature of interactions between MWE processing and downstream applications. They proposed a conceptual framework within which challenges and research contributions can be positioned. The work offers a shared understanding of what is meant by "MWE processing," distinguishing the subtasks of MWE discovery and identification.

Research questions

The current study poses the following research questions: -

1. Are there dialect-specific MWEs?
2. How can the semantic idiosyncrasies of these MWEs be analysed?
3. How can the MWEs be classified?

Aim

The current study aims to identify the semantic idiosyncrasies of selected MWEs from the Namboothiri dialect and to classify them into various categories.

Objectives

1. To compile the semantic idiosyncrasies MWEs from Namboothiri dialect.
2. To analyse the semantic idiosyncrasies of MWEs from Namboothiri dialect.
3. To classify them into various categories of MWEs.

Research Methodology

The current study, which undertakes an indirect approach, employs qualitative research design techniques in the data collection and analysis process.

Data Collection

The data required for the current study were collected using a random sampling method from primary as well as secondary sources.

Multiword Expressions (MWEs)

A multiword expression (MWE) consists of two or more words that function together as a single unit of meaning. Notably, the characteristics of an MWE cannot be anticipated from the individual properties of its component words, and its overall meaning is often distinct from the meanings of those components when considered separately

Definitions of MWEs

According to Seretan (2011) there are a lot of definitions for MWEs. Some of the important ones are the following: -

"a multiword unit or a collocation of words that co-occur together statistically more than chance" (Carpuat and Diab 2010)

"a sequence of words that acts as a single unit at some level of linguistic analysis" (Calzolari et al. 2002)

"Lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity" (Baldwin and Kim 2010)

Properties of MWEs

1. Non-compositionality and non-literal translatability

Non-compositionality is a striking feature of MWEs. This feature highlights that the meaning of the expression as a whole, the MWE, cannot be inferred from its individual components. It has its own semantic idiosyncrasy. Hence, translation of MWEs on a word-by-word basis leads to unnatural, ungrammatical, and nonsensical results. In fact, the non-compositionality feature

complements the non-literal translatability feature. Multiword expressions are idiomatic by definition. But this feature is not observed throughout all kinds of multiwords. Multiword expressions can have compositional or non-compositional semantics.

2. Ambiguity

MWEs are ambiguous when its compositional words can co-occur without forming an expression. These expressions can act as an MWE or can take the literal meaning of the sequence. This selection is contextual

3. Discontinuity

Parts of certain MWEs may get separated from each other by some external element/s. Depending on the context the intervening word may change. This makes it difficult to identify multiword expressions from a sentence.

4. Non-substitutability

Non-substitutability is a property that is relevant for most MWEs. According to this property, it is not possible to replace a part of an expression with a synonym or similar word. It often causes lexical rigidity.

5. Frequency and collocation

One of the typical properties of MWEs is that the constituent words tend to occur (together) more than expected. When compared to the chances of using a possible alternative, the frequency of co-occurrence of the component words of an MWE is larger. Since the language speakers tend to use MWEs instead of explaining the concept, multiwords happen to occur frequently. Multiword expressions are stored in the mental lexicon of language speakers. They become habitual through frequent usage. Frequency can be considered as a reliable criterion for lexicalization, but it should not be a necessary one.

Single lexical unit:

Multiword expressions consist of a minimum of two words that cut across word boundaries and are complex than the individual units. Generally, MWEs do not cross the sentence boundaries and are treated as single lexical units. The component words do not act individually.

Instead, they work together as a group and contribute meaning to the expression as a whole. They are stored as a single unit or a particular concept in the mental lexicon of the speaker.

Syntactic fixedness:

MWEs are considered syntactically fixed expressions. However, in the opinion of many linguists, MWEs exhibit a continuum of syntactic fixedness.

Spelling:

MWEs are widely seen as words with spaces. Defining multiword expressions as words with spaces is theoretically unsatisfactory. The speakers are not accurate all the time and spelling are not always consistent.

Classification of MWEs

MWEs are classified into lexicalized phrases and institutional phrases. Lexicalized phrases "have at least in part idiosyncratic syntax or pragmatics". Lexicalized phrases are again classified as fixed, semi-fixed and syntactically flexible expressions. (Sag et al., 2002)

Fixed Expressions

Fixed expressions are frozen expressions that do not undergo any morphosyntactic variations or internal modifications. They can be considered as words-with spaces. Generally, they are transparent in meaning.

Semi-Fixed Expressions

In the case of semi-fixed expressions, word order and composition are strictly invariable. However, some lexical variations are possible. Semi-fixed expressions can be further classified into three subcategories:

Non-decomposable idioms: We cannot analyze or understand non-decomposable idioms from the words they are composed of. They are semantically opaque and do not undergo syntactic variability. But they can take inflections and reflexive form variations.

Compound Nominals: Compound nominals do not undergo syntactic variations. But they do inflect for number.

Proper Names/Named entities: Proper names are syntactically highly idiosyncratic in nature.

Syntactically Flexible Expressions

Unlike semi-fixed expressions or fixed expressions, syntactically flexible expressions allow a range of syntactic variations. Syntactically flexible expressions include:

Verb-Particle Constructions: These are the expressions that consist of a verb and one or more particles and they can be compositional or semantically idiosyncratic.

Decomposable Idioms: Decomposable idioms are syntactically flexible to some extent. It is very difficult to predict the syntactic variations they undergo.

Light Verbs: Light-verb constructions are highly idiosyncratic. They undergo full syntactic variability.

Institutionalized Phrases

Institutionalized phrases are syntactically and semantically compositional, but statistically idiosyncratic. They occur with high frequency and undergo full syntactic variability.

Malayalam and its various dialects

Malayalam, the native language of Kerala in southern India, is spoken by over 35 million people and is recognized as one of the classical and official languages of India. Malayalis, the people who speak Malayalam, derive their name and collective identity from the language itself. As a member of the South-Dravidian branch of the Dravidian language family, Malayalam is considered relatively young among its linguistic relatives.

The language exhibits numerous dialects, reflecting variations in pronunciation, vocabulary, grammar, and phonology based on region, community, occupation, social status, and communicative context. These dialects are broadly categorized as regional and social. Clear distinctions exist among Southern, Central, and Northern regional dialects. Social stratification is less obvious, except for the distinct speech patterns of the Muslim community. Christian and Hindu dialects are more similar to each other than to the Muslim dialects, and caste-based dialect differences can be observed in kinship terms as well.

Research from a Kerala village (Subramoniam, 1977) reveals that:

- The Brahmin dialect is closest to the standard version of Malayalam.

- Nair dialect aligns with Brahmin speech, while the Ezhava-Tiyya dialect is closer to the Harijan dialect.
- Harijan dialects retain archaic features like pronominal endings.
- Nair and Ezhava dialects are more similar to each other than to the Muslim dialect.
- Higher-caste groups often maintain non-native phonology, whereas lower castes tend to adapt loanwords to native phonology.

Namboothiri Dialect

Namboothiris have a distinct dialect that is so much a part of their culture. As in any other speech form, culture has its overwhelming influence over the Namboothiri dialect too. They also have a peculiar intonation, diction, and style of delivery in ordinary conversation. (Nambudripad, K.D.2008) The Namboothiri dialect is a distinctive sociolect of Malayalam. It stands out for its unique blend of Sanskrit influences and Dravidian linguistic features, making it quite different from other Malayalam dialects spoken in the region (Calicut, n.d.).

Historical and Cultural Roots

The insular and priestly nature of the Namboothiri community shaped the dialect over centuries, intertwining language use with rituals, social hierarchy, and identity. This dialect evolved with heavy borrowings from Sanskrit due to the community's role in Vedic traditions and ritual recitation. Simultaneously, local Dravidian phonology contributed to the dialect's unique sound system.

Phonological Features

The Namboothiri dialect is noted for:

- Aspirated consonants (like /ph/ in /phalam/ and /dh/ in /dhanam/), which are pronounced distinctly by older speakers—this is rare in casual speech among other Malayalam speakers.
- A slightly expanded inventory of consonant clusters and nasal sounds compared to standard Malayalam.

Lexical and Syntactic Characteristics

- There are marked differences in how speakers address equals and those of different castes; for example, verbs and responses have unique endings or forms. (Calicut, n.d.-a)

Scope of the study

- The study can be further extended to all aspects of idiosyncrasies namely statistical, syntactic, phonological etc. of MWEs. It can be used in NLP based projects and studies. The study can be modelled to conduct similar studies in the dialects of other languages.
- One promising direction is the statistical analysis of MWEs, which involves examining the frequency, distribution, and co-occurrence patterns of MWEs within natural language corpora. Understanding these statistical properties can help reveal which expressions are more entrenched or culturally salient among speakers, providing deeper insights into language use and variation.
- Syntactic analysis is another vital area to extend research. This involves investigating how MWEs behave within sentence structures, including their syntactic flexibility or fixedness, the way they integrate with surrounding words, and any unique grammatical patterns they exhibit. Syntactic study can uncover rules governing the position and form of these expressions in discourse, clarifying the interplay between structure and meaning.
- Phonological aspects are equally important, particularly in a dialect marked by distinctive pronunciation and intonation patterns like Namboothiri Malayalam. Research can focus on the sound patterns within MWEs, such as phoneme realization, stress, nasalization, or aspirated consonants, which may differ from standard Malayalam or other regional dialects. These phonetic features often carry social and identity markers, making their study crucial for a holistic understanding.
- Besides theoretical linguistics, these expanded studies hold great potential for applications in Natural Language Processing (NLP). Incorporating dialect-specific MWEs into NLP models can improve machine translation accuracy, speech recognition systems, sentiment analysis, and text mining for Malayalam, especially for dialectal variants. It can enable more culturally sensitive and precise computational tools that respect linguistic diversity.

- Finally, the methodologies and findings from the Namboothiri dialect study can serve as a valuable blueprint for investigating MWEs in other dialects and languages. Comparative dialectology can benefit from such frameworks, facilitating cross-linguistic studies that enhance understanding of language contact, change, and cultural identity across linguistic communities.

Data Analysis

The analysis of the data excerpts, MWEs from the Namboothiri dialect is as follows: -. The MWE *akalōm n̲īl̤ōm koḷḷikuka* carrying the whole meaning 'make things difficult' when broken down into its constituent meanings reflect as *akalōm* means 'distance', *n̲īl̤ōm* means 'length' and *koḷḷikuka* means 'to contain' which starkly highlight its non-compositionality. This expression falls into the category of non-decomposable idiom as it cannot be analysed or understood from the words they are composed of. They are semantically opaque and do not undergo syntactic variability. Similarly, the MWEs *aḍukkaḷa māruka* with whole meaning 'to get married' and constituent meanings-*aḍukkaḷa* means 'kitchen' and *māruka* means 'change', *ambalakkāḷa* with the whole meaning 'one who leads life in a dareless manner' and constituent meanings-*ambalam* means 'temple' and *kāḷa* means 'bull' , *ambalam viḻuṅuka with whole meaning* 'to loot completely' and constituent meanings- *ambalam* means 'temple' and *viḻuṅuka* means 'to swallow', *ōtikkōn parayun̲n̲atellām ōttalla* with the whole meaning 'it is not necessary that everything said by an expert or master is always right' and constituent meanings-*ōtikkōn parayun̲n̲atellām* means 'all that is said by the expert or master' *ōttalla* means 'it is not the Veda', *ōlanmaṭṭə* with the whole meaning 'without any decorations', in other words, 'in a simple and carefree way' and constituent meanings- *ōlan* is 'a vegetarian dish that is spice less' and *maṭṭə* means 'nature of a thing', *kāppuketṭuka* with the whole meaning 'to get married' and constituent menaings- *Kāppu* means 'sacred thread tied around the wrist of the bridegroom for safety' *keṭṭuka* means 'to tie' , *kumiḻ pole* with the whole meaning 'a matter or something that lacks depth' and constituent meanings- *kumiḻ* means 'the gamhar tree'*pole* means 'like', *kuḷi bandham* with the whole meaning 'not even the slightest trace of any relationship' and constituent meanings- *kuḷi* means 'bath' and *bandham* means 'relationship', *koṭṭakuḍayukaa* with the whole meaning 'destruction of everything that is in possession' and constituent meanings- *koṭṭa* means 'basket' *kuḍayukaa* means 'to empty it' and *gōpiyākkuka* with the whole meaning 'to defeat' and constituent meanings- *gōpi* is a term carrying different

connotations depending on the context such as ‘a proper name’, ‘female counterpart of a person who looks after cattle’ etc. *ākkuka* means ‘to make’, all fall into the category of non-decomposable idioms due to their non compositionality feature.

Another category of MWE found out from the analysis of the data is the code-mixed expressions. MWEs like *uruḷakkupastarikuka* with the whole meaning ‘spend on things separately’ and constituent meanings- *uruḷa* means something that is made into a round or sphere shape, e.g, rice balls *upastarikuka* means ‘to serve ghee’ , *kaṇḍḍhasūtram parikuka* with the whole meaning ‘a Namboothiri woman becoming a widow’ and constituent meanings- *kaṇḍḍhasūtram* means ‘the marriage knote around the bride’s neck’ *parikuka* means ‘to pluck’, *kōritarppikuka* with the whole meaning ‘to serve without any limit or to spend without any limit’ and constituent meanings- *kōri* means to ‘take handfull’ and *tarppikuka* means ‘to carry out the ritual for pleasing god, departed souls etc.’ and *gaṇapati kurikuka* with the whole meaning ‘to mark the beginning of something’ and constituent meanings- *gaṇapati* ‘a Hindu god’ *kurikuka* means ‘to scribble’ belong to this category.

Discussion and Findings

The study conducted on selected Multi-Word Expressions (MWEs) from the Namboothiri dialect of Malayalam highlights the prominent feature of non-compositionality. In almost all the MWEs analyzed, the meanings of the individual constituents did not contribute directly to the whole expression. For example, in *ambalam viḻuṅuka*, which means "to loot completely," the separate meanings—*ambalam*, "temple," and *viḻuṅuka*, "to swallow"—do not combine to reveal the actual idiomatic meaning. This characteristic of non-compositionality also leads to non-translatability; accurately translating these MWEs into another language is challenging, as direct translation would often fail to carry over their exact sense and cultural resonance.

Another significant feature observed is the property of non-substitutability. The constituents of these MWEs generally cannot be replaced with other lexical items without altering or destroying their established meaning, making them lexically rigid. For instance, with the expression *kuḷi bandham*, which means "not even the slightest trace of any relationship," the parts cannot be switched out for synonyms without losing the idiomatic meaning entirely.

Collocations were also frequently noticed among these MWEs. For example, in *ōtikkōn parayuṉṉatellām ōttalla,* meaning "it is not necessary that everything said by an expert or

master is always right," the elements *ōtikkōn*, "expert/master" and *ōtt,* "Veda" form a habitual collocation.

The MWEs examined in this study were found to fall into two categories: non-decomposable idioms, and code-mixed MWEs. By illuminating these features, the analysis enhances our understanding of how MWEs in the Namboothiri dialect encapsulate cultural knowledge, reinforce lexical rigidity, and embody unique semantic identities.

Conclusion

MWEs in the Namboothiri dialect of Malayalam language exhibits semantic idiosyncrasies that contributes to their uniqueness. This property of these MWEs thus forms the basis of their classification into various categories. Unlike ordinary phrases, these MWEs cannot be understood simply by analyzing the literal meanings of their individual components. Instead, their meanings emerge from culturally and socially grounded interpretations that are specific to the Namboothiri community. This idiosyncrasy plays a crucial role in classifying MWEs into various categories, such as institutionalised phrases, non-decomposable idioms, and code-mixed expressions. These classifications help linguists and researchers organize the complex variety of expressions, clarifying how language operates as a marker of social identity and cultural heritage within the community.

This classification is not merely linguistic but also deeply anthropological. It reflects how the Namboothiri dialect encodes traditional values, rituals, and social hierarchies through language. For example, certain phrases are embedded in ritual practices and social norms that define group membership and interpersonal relations. By studying these MWEs, one gains insight into how language functions not only as communication but also as a living repository of collective history and social practices. This underscores the essential role of dialectal variation in preserving cultural identity and highlights the dynamic interaction between language and society.

Thus, the study's emphasis on semantic idiosyncrasy enriches the understanding of the Namboothiri dialect's complexity, showing how linguistic forms carry intertwined cultural meanings. This knowledge supports both theoretical linguistics and practical applications, enabling a deeper appreciation of the dialect's role in shaping and reflecting the identity of its speakers.

References

Carpuat, M., & Diab, M. (2010). *Task-based evaluation of multiword expressions: A pilot study in statistical machine translation*. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)* (pp. 242–245). Association for Computational Linguistics.Clark, A., Fox, C., & Lappin, S. (2010). *The Handbook of Computational Linguistics and Natural Language Processing*. http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number=025133212&sequence=000001&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA

Cyriac Anjali Treesa, Devi Shobitha Lalitha (2022). Classification of Multiword Expressions in Malayalam. *Proceedings of the WILDRE-6 Workshop @LREC2020*, pages 55–59

Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword Expression Processing: A survey. *Computational Linguistics*, *43*(4), 837–892. https://doi.org/10.1162/coli_a_00302

Jayaraj, T., & Kumar, R. (2019). Variation of Lexical Items and the Changing Need of Identity Projection: A Study on Mappila Malayalam. Dialectologia. https://doi.org/10.1344/dialectologia2019.24.7

Jisha, T.E. & Thomas Monoth (2017). Identification of Multiword Expressions: A Literature Study.

Kumar Shivek, Behera Pitambar, Jha Nath Girish (2017) A classification-based approach to the identification of Multiword Expressions (MWEs) in Magahi Applying SVM Namboodiripad, S., & Garellek, M. (2017). Malayalam (Namboodiri Dialect). Journal of the International Phonetic Association, 47(1), 109–118. doi:10.1017/S0025100315000407

Padmanabhapillai. (2019). Sabdatharavali. In *Sabdatharavali* (44th ed.). Sahitypravarthaka Sahakarana Sangham.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In A. F. Gelbukh (Ed.), *Computational linguistics and intelligent text processing: Proceedings of the Third International Conference on Computational Processing of Portuguese (PROPOR'2002)* (Vol. 239, pp. 1–15). Springer.

Seretan, V. (2011). *Syntax-based collocation and multiword expression detection*. Morgan & Claypool Publishers.

Singh, D., Bhingardive, S., & Bhattacharyya, P. (2016). Multiword Expressions Dataset for Indian Languages. *Language Resources and Evaluation*, 2331–2335. https://dblp.uni-trier.de/db/conf/lrec/lrec2016.html#SinghBB16

Survey: Multiword Expression Processing: A Survey (Constant et al., CL 2017)

Vasudevan,Jayan, Bhadran, V.K., Raghunathan,Arathi (2016) Contextualizing Multi-Word Expressions in English and Malayalam

Vishnu Namboothiri, M.V. (2017). Namboothiri Bhasha Nighandu. Kerala Bhasha Institute

Calicut, N. W. (n.d.). *Namboothiri dialect*. https://www.namboothiri.com/articles/dialect.htm