Proceedings of the WILDRE-6 Workshop @LREC2020, pages 55–59
Marseille, 20 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Classification of Multiword Expressions in Malayalam

Treesa Anjaly Cyriac, Sobha Lalitha Devi
AU-KBC Research Centre, MIT Campus of Anna University, Chennai, India
treesaanjaly07@gmail.com, sobha@au-kbc.org

Abstract

Multiword expression is an interesting concept in languages and the MWEs of a language are not easy for a non-native speaker to understand. It includes lexicalized phrases, idioms, collocations etc. Data on multiwords are helpful in language processing. 'Multiword expressions in Malayalam' is a less studied area. The boundary between multiword expressions and other compositions is fuzzy. Not all the multiword expressions adhere to all the properties of MWEs. In this paper, we are trying to explore multiwords in Malayalam and to classify them as per the three idiosyncrasies: semantic idiosyncrasy, syntactic idiosyncrasy, and statistic idiosyncrasy. Though these are already identified, they are not being studied in Malayalam. The classification and features are given and are studied using Malayalam multiwords. Through this study, we identified how the linguistic features of Malayalam such as agglutination influence 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.

Keywords: Multiword expressions, NLP applications, idioms, Malayalam, Dravidian language, linguistic idiosyncrasy, MWE, Machine Translation, lexicalized expressions

1. Introduction

According to Sag et al. (2002), multiword expressions are "idiosyncratic interpretations that cross word boundaries". They show semantic, statistic, and syntactic idiosyncrasy. Multiword expressions are word sequences that act as a single lexical unit. The meaning of the individual components does not contribute to the collective meaning of the expression. It is difficult for humans to understand the underlying meaning of such expressions. It is even more difficult for a machine to resolve these expressions. To tackle this problem we need more linguistic analysis of multiwords. Machine translation has been helpful for language learners, non-native speakers, and even translators. Solving linguistic barriers could be the beginning of productive collaborations and innovations. But using machine translation systems to solve the problem of MWEs is not fruitful often because they lack good input regarding multiwords. Available information about Malayalam MWEs is insufficient to come up with an effective translation system that addresses this linguistic concept which Sag. et al. called 'a pain in the neck of NLP'. Through this research, we are trying to study the properties and types of Malayalam MWEs which would help in improving machine translation systems.

Multiword disambiguation is very important for language processing. Most of the time, the translation system gives a literal translation of the individual words. For example, consider the Malayalam word (kathakalikkuka | lit. 'To end the story.') 'To kill/ to end/ to defeat.' If the input is 'kathakalikkuka' Google translate translates it as 'eat the story'. If the input is given without a space in between i.e. 'katha kalikkuka system translates it as 'tell the story'.

The paper is divided into seven main sections: The first section is the Introduction. The second section deals with relevant previous works which is followed by the Classification of MWEs. In the fourth section we discuss about the Types of Multiword Expressions and the Properties of MWEs in the fifth section. Findings are mentioned in the sixth section and the Conclusion in the seventh section. The examples used to substantiate the classification, types and properties are randomly taken from the language.

2. Related Work

We explored many previous studies and in this section we are trying to explain how their findings helped the present work.

Ivan Sag et al. (2002) classify MWEs into lexicalized phrases and institutionalized phrases and it gives further classification for lexicalized phrases. The paper also provided some analytic techniques for MWEs. They used the constraint based Head-driven Phrase Structure Grammar formalism. Rules and disambiguation strategies in the English- Malayalam Machine Aided Translation system (AnglaMalayalam) has been discussed in Vasudevan et al. (2016). According to the authors, the English- Malayalam Machine Aided Translation system based on AnglaBharati Technology which is discussed in this paper showed good results after introducing these rules.

Lahari Poddar (2016) presents some of the features and classifications of multiword expressions and different approaches towards their automatic extraction. The paper also presents numerous examples from Indian languages. Tanmoy Chakraborty (2011) presents a vast study on multiwords with main focus on Bengali MWEs. This paper also presents different types and properties of MWEs. The paper gives classification of MWEs in Bengali. The study modeled the syntax and semantics of Bengali MWEs based on the statistical approaches of substitutability, co-occurrence properties, semantic clustering and linguistic properties. Timothy Baldwin and Su Nam Kim (2010) shed light to the research issues relating to MWEs.

3. Classification of MWEs

Sag et al. (2002) classifies MWEs into lexicalized phrases and institutionalized phrases. Many other classifications come under these broad terms.

3.1 Lexicalized Phrases

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

3.1.1 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.

3.1.2 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. Examples: (kuḷam thoṇṭuka| lit. 'To dig the pond.') 'To destroy.'

Compound Nominals: Compound nominals do not undergo syntactic variations. But they do inflect for number. Compound nominals such as (erivumpuḷiyum | lit. 'spicinessand sourness') 'Taste' are very frequent.

Proper Names/Named entities: Proper names aresyntactically highly idiosyncraticin nature.

1) mahatmagāndhi saṛvakalāśāla
   Mahatma Gandhi University
2) trśūṛ pūram
   Thrissur Pooram
   A temple festival held in the district of Thrissur.

3.1.3 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.

3) kaḷaňňu kuḷikkuka
   wasted bathe
   lit. 'To waste and bath.'| 'To fritter.'
4) peṭṭpōvuka
   happen to go
   lit. 'Get into.' | 'To get trapped.'

Decomposable Idioms: Decomposable idioms are syntactically flexible to some extent. It is very difficult to predict the syntactic variations they undergo. (mūkkumkuttivīḻuka |lit. 'To fall upside-down.') 'Plummet.' is a decomposable idiom.

Light Verbs: Light-verb constructions are highly idiosyncratic. They undergo full syntactic variability. Expressions like (tīrumānam eṭukkuka | 'Take a decision.') are light verb constructions.

3.2 Institutionalized Phrases

Institutionalized phrases are syntactically and semantically compositional, but statistically idiosyncratic. They occur with high frequency and undergo full syntactic variability. Phrases such as (erivum puḻiyum| lit. 'spiciness and sourness') and (tekk vaṭakk naṭakkuka | lit.'walk south to north.') 'Walk/ live aimlessly.' are institutionalized phrases.

4. Types of Multiword Expressions

Multi-word expressions can be grouped into the following types: Reduplication, partial reduplication, semantic relationship, code-mixed multiwords, collocations and compound verbs.

4.1 Reduplication

Reduplication is a word-formation process by which the root or stem of a word, or a part of it, is repeated to produce meaning. Exampls : ḍumḍum (knoking sound), ōṭiyōṭi (ran continuously), jillampaṭapaṭa (the sound of a musical instrument, chenda), payyepayye (slowly) etc.

4.2 Partial Reduplication

In partial reduplication, the given word is partially replicated. Examples: vīṭvīṭāntaram (house to house), talaṅṅumvilaṅṅum (hither and thither) etc.

4.3 Semantic Relationship

There are expressions with some kind of semantic relationship existing between the constituent words.

Synonym: (sambalsamrddhi | lit. 'riches and abundance') 'prosperity', (āyuṛārōgyam | 'life and health') 'welfare' etc.

Antonym: (jīvan maraṇa) 'life or death situation', (dinarātṛaṅṅaḷ) 'days and nights', (sukhadukham) 'happiness and sadness' etc.

Sister Words: (veḷivum veḷḷiyāḻccayum | 'Sense and Friday')sanity', (bellum bṛeykkum | 'bell and breake') 'control', (kaṇṇum mūkkum | 'eye and nose') 'sense' etc.

4.4 Code-mixed Multiwords

Code mixed multiwords are very common in Malayalam. They are not just large in number, they occur very frequently too. Examples:

5) āyuṛārōgyam
   lifehealth
   lit. 'Life and health' | 'Welfare'
6) fyūspōyi
   fuseleft
   lit. 'The fuse tripped.' | 'Lost one's mind.'
7) tyūb-laiṛṛāyirikkuka
   tube-light be
   lit. 'To be a tube-light.'| 'To be obtuse.'
8) kḻikkākuka
   click to get
   lit.'Happen to click.' | 'To understand/ get liked by others/ to become successful.

In example (5), the word (āyuṛ | lit. 'life') is taken from Sanskrit. The first parts of (6), (7), and (8)are English words. Code-mixed multiwords are very often used

among Malayalam speakers. Some of them can have multiple meanings. For instance, example (8) can be used in different contexts:

a. enikk onnum kl̲ikk āyilla
   I anything click didn't happen
   lit. 'Nothing clicked for me.' | 'I didn't understand anything.'

b. putiya kaṭa kl̲ikk āyi
   new shop click became
   lit. 'The new shop became click.' | 'The new shop became successful.'

4.5 Collocations

Collocations are word sequences that co-occur more often than would be expected by chance. Examples: (śuddhavāyu) 'fresh air', (iṭiyumminnalum) 'thunder and lightning', (vaṇṭiyumvaḷḷavum) 'transportation' etc.

4.6 Compound Verb

A compound verb is a series of words that acts as a single verb. One part of the sequence is a light verb that can take inflections of tense, mood, or aspect. The other part carries most of the semantics and hence the key. Examples: (*ōṭṭipōyi*), 'ran away', (*kēṭṭuninnu*) 'listened without responing', (*vann kaṇṭu*) 'visit', (*uṛaṅṅi pōyi*) 'fell asleep' etc.

5. Properties of MWEs

Non-compositionality and Non-literal translatability: Multiword expressions are semantically idiosyncratic. The meaning of the whole expression cannot be inferred from the meanings of its parts. Therefore word-for-word translation tends to generate unnatural, ungrammatical and, sometimes nonsensical results.

9) cukkān piṭikkuka
   helm hold
   lit. 'Hold the helm' | 'Take the helm.'
10) kaṭakkal kōṭāli vakkuka
   At the root axe lay
   lit. 'Lay axe at the root.' | 'Put at stake'

Non-compositionality is considered a prominent feature of multiword expressions. This also compliments the feature, non-literal translatability**.** Multiword expressions are idiomatic by definition. But this feature is not observed throughout all kinds of multiwords. Consider the expressions like 'iṭiyum minnalum' (thunder and lightning), 'vaṇṭiyum vaḷḷavum' (transportation) etc. Here, the constituent words bear a direct relation to the meaning of the expressions. Multiword expressions can have compositional or non-compositional semantics. Many MWEs, especially some collocations, do not stick to this property.

Ambiguity: An MWE is ambiguous when its compositional words can co-occur without forming an expression. Example: (kaikōṛkkuka | lit. 'Join hands.') 'Work together/ collaborate.', (gyāstīruka | lit. 'Run out of gas.') 'Getting tired.'etc. These expressions can act as an MWE or can take the literal meaning of the sequence. This selection is contextual.

Discontinuity: Parts of certain MWEs may get separated from each other by a/ some external element/s. Depending on the context the intervening word may change. This makes it difficult to identify multiword expressions from a sentence. For example, the expression (paṇipāḷi | lit. 'work slipped') 'Messed up' can occur as (paṇi pinneyum pāḷi | lit. 'work slipped again') 'Messed up again'. Another example is (gyāstīruka | lit. 'Run out of gas.') 'Getting tired.'.

11) gyās mu̲luvanum tīruka
   gas completely run out
   lit.'Completely run out of gas.' | 'Exhausted.'

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. Examples:

12) kāṭ kayaṛi
   forest climbed
   lit. 'Went to the forest.'| 'To do something too much.'
   *vanam kayaṛi
   forest climbed
13) kaṇṇ maňňaḷikkuka
   eye turn yellow
   lit. 'Eye turn yellow.' | 'Lose sight under the influence of something exciting.'
   *nayanammaňňaḷikkuka
   eye turn yellow
14) uppum muḷakum
   saltchilli
   lit. 'Salt and chilli.' | 'Taste'
15) erivum puḷiyum
   spiciness sourness
   lit. 'spiciness and sourness' | 'Taste.'
16) uppum puḷiyum
   salt sour
   lit. 'salt and sourness' | 'Taste.'
17) arakkainōkkuka
   half hand try
   lit. 'Try half hand.' | 'To give something a try.
18) orukainōkkuka
   one hand try
   lit. 'Try one hand.' | To give something a try.

(mukham maňňaḷikkuka | lit. 'Face turn pale') 'Feel embarrassed' is an error-free multiword we get by substituting one item of the expression (13) with another. Here, the meaning of the expression changes. In the case of (14) and (15), both the expressions represent the same concept. But here, the substitution of a part of the former by a part of the latter can happen. I.e. (16).Though it is less frequent than the others, it still conveys the same meaning. Similarly, (17) & (18) refer to the same concept. In these examples, the final part stays constant. The initial parts, *arakkai* & *orukai*, can be used interchangeably without causing any change in meaning. Non-substitutability is not a mandatory property multiword expressions should follow.

Frequency & 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. Examples: (kīḻmēl maṛiñňu | 'Fell bottom-up.') 'Turn upside down.', (kaṇṇil eṇṇayoḻiccirikkuka | lit. 'Poured oil in the eye.') 'Wait impatiently.',(bellum bṛeykkum | lit. 'Bell and brake.') 'control' etc.

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. Consider the following expression:

19) kōl oṭikkuka
    stick to break
    lit. 'Break the stick.' | 'To give up'.

This expression is rarely used in the language. And it seems to be a regional usage. Even though multiword expressions appear to be frequent in the language they do not adhere to the property of frequency.

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 is not always consistent. Since Malayalam is an agglutinative language, there is a tendency to join words very often. For example,

20) tēcc oṭṭikkuka
    iron out to paste
    lit. 'Iron out and paste.' | 'To cheat.'
21) paṇi pāḷi
    work slipped
    lit. 'Work slipped.' | 'Messed up.'
22) kaṭiňňāṇ iṭuka
    bridle to put
    lit. 'Put a bridle.' | 'Bring under control.',
23) kai kaṭattuka
    hand to insert
    lit. 'Insert the hand.' | 'To interfere'

Examples (20), (21), (22), and (23) can be written as tēccoṭṭikkuka, paṇipāḷi, kaṭiňňāṇiṭuka, and kaikaṭattuka respectively. The native speakers show a tendency to pronounce them as a single word. This trend is seen in both written and spoken forms. This property makes them look like a compound word. Therefore, we struggle to define an explicit boundary between multiword expressions and compound words.

Unlike most compounds, we can insert external words (property of discontinuity) within some MWEs that look like compound words. For instance, (20) can be modified as (tēcc bhittiyil oṭṭikkuka | lit. 'Iron out and paste on the wall.') 'to cheat brutally'. But this is not a generic criterion. In the opinion of Bauer (2019), compounds are one type of MWE and since they overlap with other MWEs, it's not easy to define compounds.

6. Findings

From the studies we arrive at the following findings that make Malayalam multiwords different:

Agglutination: Multiwords may get agglutinated with the neighboring words or with the component words of the same expression itself.

Two-way rendering: Multiwords can be written together or separate, without any meaning change.

Code-mixed multiwords: Malayalam has a large set of code-mixed multiwords and many of them are high-frequency words.

7. Conclusion& Future Work

Green et al. (2011) point out that "MWE knowledge is useful, but MWEs are hard to identify." Types of word combinations lie in a spectrum. The boundary between multiword expressions and other compositions is fuzzy. Not all the multiword expressions adhere to all the properties of MWEs. The examples were given in this paper are randomly taken from the language.

Agglutination and two-way rendering of Malayalam multiwords are serious problems that require special attention. This information is very important for speech recognition systems to understand the dialogues by a native speaker.

Processing of multiword expressions requires contextual information. Otherwise, problems related to discontinuation and ambiguity could not be resolved.

Available data is very insufficient for the improvisation of translation systems and other NLP areas. Our future work includes preparing a glossary of Malayalam multiword expressions with wide coverage and sufficient linguistic knowledge.

Implementing computational models is also our future concern. Incorporating them in machine translation and other NLP areas could help the betterment of the system significantly.

8. References

Baldwin, T. and Kim, S. N. 2010. Multiword expressions, In N. Indurkhya and F. Damerau, (Eds.), *Handbook of Natural Language Processing, Second Edition*. CRC Press, Boca Raton, pages, 267-292.

Barreiro, A., Monti, J., Orliac, B., and Batista, F. 2013. When Multiwords go bad in machine translation. InMonti, J., Mitkov, R., Pastor, G. and Seretan, V., (Eds.), *Proceedings of the MT Summit Workshop Proceedings on Multi-word Units in Machine Trans-*

lation and Translation Technology, The European Association for Machine Translation, pages 26–33.

Bauer, L. 2019. Compounds and multi-word expressions in English. In B. Schlücker, editor, Complex lexical units. Berlin, Boston: De Gruyter, pages 45-68.

Chakraborty, T. 2011. Multiword Expressions. Master's thesis, Jadavpur University, May.

Hüning, M. and Schlücker, B. Multi-word expressions. 2015. Muller, Peter O., Ohnheiser, I., Olsen, S., and Rainer, (Eds.), Word Formation, An International Handbook of the Languages of Europe. Berlin: De Gruyter.

Kuiper, K. 2018. Multiword expressions and the Law of Exceptions. In M., Sailer, and S., Markantonatou, editors, *Multiword expressions*: *Insights from a multi-lingual perspective*. Berlin: Language Science Press, pages 121-141.

Poddar, L. 2016. Multilingual Multiword Expressions. Master's thesis, Indian Institute of Technology, Bombay.arXiv:1612.00246.

Rayson, P., Piao, S., Sharoff, S., Evert, S., and Moirón, B. 2010. Multiword expressions: hard going or plain sailing?. Lang Resources & Evaluation 44, 1-5.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Dan Flickinger.2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, 1-15. Mexico City.

Vasudevan, J., Bhadran, V.K., and Raghunathan, A. 2016. Contextualizing multi-word expressions in English and Malayalam. *Proceedings of the International conference on Dravidian linguistics*, Hyderabad.