# Indian Language MultiWord Expression Annotation Guidelines

**Version: 1.1**

Date: 11-03-2025

Prepared by

IIIT-Hyderabad

Table of Contents

**What is Multi-word Expression (MWE)**......3

Compositionality and Non-compositionality......3

Principle of Compositionality:......4

**1. Noun Compound**......5

Writing convention...... 6

1.1 When not to be considered as NC......7

**2. Reduplicated Expression**......8

**3. Echo-word**......9

**4. Opaque**......10

**5. Opaque- Idioms**......10

Note:......11

## What is Multi-word Expression (MWE)

MWEs are complex linguistic forms consisting of several words (or constituents) but conveying one meaning. We will annotate the following types of MWEs-

- Noun Compound
- Reduplicated Expression
- Echo-word
- Opaque
- Opaque-Idiom

MWEs can convey either

(i) literal meaning composed of the meaning of the constituents or

(ii) opaque meaning, or

(iii) Idiomatic meaning (which is also termed as non-compositional) meaning

### Compositionality and Non-compositionality

Compositionality is an important characteristic of MWEs as this determines the degree of contribution of meaning and features of individual words in a whole MWE. On the other hand, non-compositionality is a challenge for NLP as the meaning translation of individual words will yield an incorrect translation for the whole string. Thus, annotating non-compositional MWEs is an important task for NLP.

- **Meaning can be compositional**

Compositionality in meaning refers to the principle that the meaning of a complex expression (such as a combination of words or a phrase) is determined by the meanings of its parts and the way those parts are combined.

**Principle of Compositionality**:

- The meaning of a whole is derived systematically from its components.
- Example: *"red apple"* → meaning comes from the meaning of *"red"* and *"apple"*, additonally how they are combined.

- **English**

The ***state government*** requested funds for emergency measures but did not close access to the route.

- **Hindi**

***vana saṁrakṣaṇa*** mēṁ adībāsiōṁ kā bhūmikā mahattavapūrṇa hai

Forest conservation tribal.PL GEN role important be.PRE

“The role of tribals is important in forest conservation.”

- **Bengali**

***banyā trāṇē*** sarakāra ēka kōṭi  ṭākā anudāna diẏēchē

flood relief government.NOM one crore rupees donation give.PRES. PERF.3

‘The government has donated 1 crore rupees for flood relief.’

- **Meaning can be non-compositional**

Non-compositionality refers to cases where the meaning of a phrase, sentence, or expression cannot be directly derived from the meanings of its individual parts and their syntactic structure. In other words, the whole meaning is not predictable from the sum of its components.

- **English**
  - The bonds are trading at just 40% of ***face value***.

- This is very important because under martial law ***kangaroo courts*** have been operating and sentencing people to death.

- **Bengali**

  tumi āmāra ***aśbaḍimba*** karabē.
  2PS.NOM 1PS.GEN horse egg do.FUT.2

  “You will do nothing to me.”

---

## 1. Noun Compound

A noun compound consists of two or more nouns with <u>the final noun head and other nouns as modifiers.</u>

- **English**

  - I want a wooden chair, not an ***iron chair***.
  - The new ***iron office chair*** is comfortable for sitting.
  - Riya is staying in a ***women hostel*** near Vijaynagar.
  - ***Women hostel campus*** needs to take all safety precautions during Holi.
  - A recent published medical bulletin claims that ***cancer death rate*** is increasing in India.

- **Hindi**

  ***kṛṣi praśāsana udyoga*** ne gāṃvoṃ meṃ kṛṣi vikāsa ko baḍha◌़◌ावā diyā

  Agriculture process industry.NOM village.LOC agriculture development.ACC promote.PAST.3.
  ‘Agricultural administration industry promoted agricultural development in villages’

rāma ***basa aḍḍe*** para hai

ram.Nom bus stop be.pres.simple.3PS
'Ram is at the bus stop.

***praveśa dvāra*** suṃdara hai

Entrance gate beautiful be.Pres
'The entrance gate is beautiful

**Bengali**

rāju ekajon ***grihasikkhak***
Raju.NOM one.CLA home tutor
'Raju is a home tutor'

- When an adjective modifies the modifier, that adjective will be part of the noun compounds
  - **Hindi**

***vanya jīva surakṣā kēṁdra*** mēṁ abhyarthiyōṁ kō praśikṣaṇa pradāna kiyā jātā hai|

Wild life protection centre.LOC candidate.PL.DAT training give.PRES.PASS.3

'Training is provided to the candidates at wildlife protection centre'

Here, *vanya* 'wild' modifies *jīva* 'animal' which in turn modifies the head *keṃdra* 'center' of the noun compound. Therefore *vanya* will be part of the noun compound while *the new iron chair* will not be part of the compound because *new* modifies the head *chair*.

**Writing convention**

Writing convention also plays an important role in Noun Compound annotation. We will follow the following schema of writing convention while annotating NC -

- **Hindi**

| | | | |
|---|---|---|---|
| AB | gṛha śikṣaka | Home tutor | rāju eka ***gṛha śikṣaka*** hai<br>Raju.Nom one home tutor be.Pres<br>'Raju is a home tutor' |
| A-B | gṛha-śikṣaka | home-tutor | rāju eka ***gṛha-śikṣaka*** hai<br>Raju.Nom one home tutor be.pres<br>'Raju is a home tutor' |

## 1.1 When not to be considered as NC

- **When written as one word in Hindi**

rāma āja ***vidyālaya*** nahīṃ gayā

ram.NOM today school neg go.PAST
‘Ram did not go to school today.’

rāju eka ***gṛhaśikṣaka*** hai

Raju.Nom one home tutor be.pres

‘Raju is a home tutor'

eka ***lohe kā ciyara*** lāo

One iron.GEN chair bring.IMP
‘Bring one chair of iron.’

- **Bengali**

ismār ***cokhera maṇi*** bādāmi

Isma.GEN eye.GEN pupil brown

‘Isma’s pupil is brown.’

jotīn ***bārudēra stūpē*** āgun chum̐ṛē diyechilo

Jotin.NOM gunpowder.GEN pile.LOC fire throw give.PAST.3

‘Jotin threw fire in a pile of gunpowder.’

- When an adjective modifies the head of the noun compound, that adjective will not be part of the noun compounds

- **English**

○ The administrative office has ordered 100 new ***iron office chairs***.

○ Our students use computer aided ***data management*** for corpus evaluation.

---

## 2. Reduplicated Expression

When a word or part of it is repeated to make a new meaning, we call it a reduplicated expression and annotate it as MWE. We consider echo-word also as a reduplicated expression if they are written with a space.

- **Hindi**

rāma ***kabhī-kabhī*** skūla ātā hai

ram.Nom sometime sometime school.Acc come.PRES.3PS
‘Ram sometimes comes to school’

- **Bengali**

***ghare ghare*** cithi geche

house. Loc house letter.Nom go.Past'
'Letters have been sent to each house'

---

## 3. Echo-word

An echo word is a linguistic term that refers to a partially repeated form of the base word, where the initial phoneme or syllable is replaced by another phoneme or another syllable. In most languages in which this phenomenon is present, echo words serve to express the meaning of "... and such; and things like that.

- **Hindi**

**caye-saye** pio
tea.ECHO drink.imp
‘(You) Have tea etc.’

- **Bengali**

**phul-tul** eno na

flower.ETC bring.imp neg

“(You) Do not bring flowers etc.’

---

## 4. Opaque

An opaque expression is  expressions or phrases whose meaning cannot be derived or predicted from the meanings of their individual constituent parts or the meaning of the whole is not simply the sum of its parts.

**Example**

**White ant-** which does not mean an ant which is white, but a type of termite.

## 5. Opaque- Idioms

Idiom is a phrase or expression that usually presents a figurative, non-literal meaning attached to the phrase.

- **Hindi**

  sabhī ne apanī ***kamara kasa*** lī thī
  everyone .NOM own waist tie take.PAST.3
  ‘Everybody was prepared.’

- **Bengali**

  jotin ājkāl ***dumurer phul*** hoye geche
  jatin.NOM nowadays fig.GEN flower be.non-F go.PRES.PERF.3
  ‘Jatin has disappeared completely nowadays.’

## Note:

We have decided not to annotate complex predicates or compound verbs as MWEs. However, if a complex predicate or compound verb functions as an opaque or idiomatic expression, we will annotate it accordingly.

**Bengali**

*rāma* ***kēṭē*** ***paṛalō***
Ram.NOM cut.Non-F fall.PAST
'Ram ran away.'

In this examples, *kēṭē paṛalō* is a compound verb (***kāṭā*** **'to cut' +** ***pōṛā*** **'to fall'**). However, the meaning derived from this verb-verb combination is unrelated to either 'cut' or 'fall.' Therefore, the annotator will classify it as 'Opaque.'

- **Hindi**
  agar tumne samay par fees nahi bhari, to college mein dakhile se
  **haath dho baithoge.**
  If 2PS.NOM time on fees neg pay.COND then college in admission from
  hand wash sit.FUT
  " If you don't pay the fees on time, you will completely lose your college admission."

- **Bengali**
  pāṛāra dādu gatakāla rātē **paṭōla tulēchēna**
  neighbourhood.GEN grandfather.NOM yesterday night pointed gourd lift.Pres.Perf
  'The neighbourhood grandfather died last night.'

In this examples, *hāth dho baithoge (**hāth*** **'hand' +** ***dho*** **'wash' +** ***baith*** **'to sit'**) and *paṭōla tōlā* is a complex predicate (***paṭōla*** **'pointed gourd' +** ***tōlā*** **'to lift'**). However, the meaning derived from these noun-verb combinations is unrelated to either 'hand' 'wash' and 'to sit' or 'pointed gourd' and 'to lift' respectively. Moreover, the opacity of the noun-verb combinations, they are used as idioms in Hindi and Bangla respectively. Therefore, the annotator will classify these as 'Opaque-Idiom'.