Indian Language MultiWord Expression Annotation Guidelines

Version: 1.1

Date: 23-03-2024

Prepared by

IIIT-Hyderabad

1. <u>What is Multi-word Expression (MWE)</u>

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
- Complex Predicate
- Compound Verbs
- Reduplicated Expression
- Idiom

MWEs can convey either

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

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

We will give examples of each case below:

2. <u>Noun Compound</u>

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 baḍha़āvā 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 ***griha sikkhak***
  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*.

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**

  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.PRES

“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**

  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.
- Hindi

- BENGALI
- tumi āmāra ***aśbaḍimba*** karabē.

  2PS.NOM 1PS.GEN horse egg do.FUT.2

  “You will do nothing to me.”

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

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

2.1 When not to be considered as NC

- **When written as one word**

  Hindi

  - rāma āja ***vidyālaya*** nahīṃ gayā
    ram.NOM today school neg go.PAST.3
    ‘Ram did not go to school today.’

  - rāju eka ***gṛhaśikṣaka*** hai
    Raju.Nom one home tutor be.pres
    'Raju is a home tutor'

- **Having any post-position or suffix in between**

  Hindi

  - 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.’

Bengali

- 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.

3. Complex Predicate

When Noun/Adj + Verb make one predicate we call it Complex Predicate and this will be annotated as MWEs.

Hindi

- rāma ***snāna kara*** rahā hai
  ram.Nom bath do.PRES.PROG.3PS
  ‘Ram is taking a bath.’

Bengali

- binā kāroNe ṭina lina-r upor ***rāg korlo***
  without reason.LOC Tina.Nom Lina-GEN upon angry do.PAST.3PS
  ‘Tina became angry with Lina without any reason.‘

Non-compositional Complex predicate

Example

Hindi

Bengali

uni āmākē galā *dhākkā dilēna*
3PS.NOM 1PS.DAT neck push give.IMP.Hon.

“He drove me away.”

4. Compound Verb

Compound verbs consist of two verbs where a light verb or semantically bleached verb occurs with the main verb

Hindi

- ○ bacce ne khilaunā ***toḍa ḍālā***
  child.Nom toy break.PST.3PS
  The child broke the toy.

- ○ soco mata, vānī eka gānā ***gā degā***
  Worry neg vani.NOM one song sing give.FUT.3PS
  ‘don’t worry, Vani will (manage to) sing one song.’

- pākhi-ṭā ***uṛē gēlo***
  Bird-CLA.Nom fly.CPM go.PAST.3PS
  ‘The bird flew away.‘

Bengali

- rākhi rikhā kē ēkṭā putula ***baniẏe dilo***
  Rakhi.Nom Rikha-DAT one-CLA doll.Acc make.CPM give.PAST.3
  ‘Rakhi made a doll for Rikha.‘

Non-compositional Compound Verb
Example
Hindi

Bengali

sē cupicupi ***kēṭē paṛalō***
3PS.NOM silent silent cut drop.PAST.3
“He ran away silently.”

5. 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'

- ***cā tā*** khāo
  tea ECHO eat.imp
  'Have tea etc.'

6. Idioms

Idiom is a phrase or expression that usually presents a figurative, non-literal meaning attached to the phrase. We will annotate idioms as MWEs.

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.CPM go.PRES.PERF.3
  ‘Jatin has disappeared completely nowadays.’