

# Vālmīki Rāmāyaṇa Annotation Guidelines

What is annotation?

Annotation is the process of highlighting task-relevant information from a text. The type of information highlighted depends on the task.

Why do we need to annotate Rāmāyaṇa?

We, ultimately, want to create a Question-Answering (QA) system from Vālmīki Rāmāyaṇa. However, QA is a semantic and complex task and depends on several smaller tasks. The automated tools available in Sanskrit are not sufficiently accurate. Therefore manual annotation is the best way to go.

Source Text

The Digital Corpus of Sanskrit (DCS) is a Sandhi-split corpus of Sanskrit texts with full morphological and lexical analysis. We use the [Vālmīki Rāmāyaṇa](#) text corpus from DCS.

Annotation Tools

We have developed two web-based annotation tools. These will be used for annotation.

1. Antarlekhaka, a web-based multi-task annotation tool which can perform several NLP annotation tasks.
2. Sangrahaaka, a web-based tool for the annotation of entities and relationships towards construction and subsequent querying of Knowledge Graphs.

## Annotation Process

Before you start annotation first visit [Ramayana Sanskrit to Tamil : The Dharmalaya Edition](#)).

### HOW TO FIND YOUR SARGA/VERSE IN THE PDF FILE

- Search in wiki text of Mahabharata or any other version available online.  
[महाभारतम् - विकिस्रोत: \(wikisource.org\)](#)
- Notice the beginning of the Sarga and find that Sarga in the PDF file and find your verse.

### BEFORE ANNOTATING

- Notice the Anvaya order in the PDF file and be familiar with the Anvaya sentence. Decide where to insert sentence boundary.
- When you compare PDF and tool verses in the Anvaya task, there might be some confusion as the tool will show many extra words because it will include Lemma words from the corpus. [example सन्धि words and सन्धि-split words, समास and members of समास words as separate words].
- We need to retain सन्धि-split words. So remove सन्धौ words. Do not reverse the phonological changes that take place in Sandhi.
- Think which words to retain and which to remove. Retain समास words as it is. If the tool is showing members of समास word, then remove such words with the help of the delete words button. Clue: In the Features section, case is marked as Cpd for such words.

NOTE: Please document the verse numbers where you have changed the Anvaya order from the Dharmalaya version.

## ANNOTATION


- Login to <https://sanskrit.iitk.ac.in/valmikiramayana/antarlekhaka>
- Select the chapter assigned to you
- Select the first verse from the chapter you are annotating

You will see an interface as shown below,

वाल्मीकिरामायणम् - Bāla 1 (part 2)   

**Boundary** Anvaya NER Coreference

Verse	Text	≡
40	वार्यमाणः सुबहुशो मारीचेन स रावणः न विरोधो बलवता क्षमो रावण तेन ते	
Word	वार्यमाणः	सुबहुशो सु
Lemma	वारय्	– सु
UPOS	VERB	– ADV
XPOS		
Features	Case=Nom Gender=Masc Number=Sing	

**Sentence Boundary**  

Previous Verse

40 वार्यमाणः सुबहुशो मारीचेन स रावणः  
न विरोधो बलवता क्षमो रावण तेन ते

41 अनाद्य तु तद् वाक्यं रावणः कालचोदितः  
जगाम सहमरीचस् तस्याश्रमपदं तदा

40 

Here, you will perform 5 tasks.

1. Sentence Boundary Detection (shown under **Boundary** tab)
2. Anvaya
3. Named Entity Recognition (NER)
4. Coreference Resolution
5. Action Graph

## Sentence Boundary Detection

A sentence in Sanskrit may span across multiple verses, or a single verse may contain multiple sentences as well. In this task, we identify the sentence boundaries and insert “ ## ” (two hash symbols surrounded by spaces) at the boundaries.

- Select the verse from the tool.
- Find your Sarga/Verse from the above resource. Read and note the Anvaya order.
- Match the verse and make sure you find all the words listed in the anvaya order of the PDF file and only then insert a Sentence Boundary marker (##).

NOTE: Please ensure that you give a blank space after a word before inserting ##, i.e. If you want to insert ## after अस्ति, the correct way to insert boundary is

अस्ति ## and not अस्ति##

- Submit.

Boundary

Anvaya

NER

Coreference

Sentence Boundary

1

तपःस्वाध्यायनिरतं तपस्वी वाग्विदां वरम्  
नारदं परिप्रच्छ वाल्मीकिर् मुनिपुंगवम् ##

2

को न् अस्मिन् साम्प्रतं लोके ## गुणवान् कश् च वीर्यवान् ##  
धर्मज्ञश् च कृतज्ञश् च सत्यवाक्यो दृढव्रतः ##

3

चारित्र्येण च को युक्तः सर्वभूतेषु को हितः  
विद्वान् कः कः समर्थश् च कश् चैकप्रियदर्शनः

2

Submit

## Word Order (Anvaya)

After marking the boundaries, words need to be arranged in the Anvaya order. Any printed edition (e.g. [Ramayana Sanskrit to Tamil : The Dharmalaya Edition](#)) may be used as a reference. Adhyāhāra tokens can also be added.

After submitting the sentence boundaries, the tool will automatically take you to Anvaya annotation interface. Here, the tokens from the sentence will be shown to you. These can be dragged using a mouse pointer to rearrange them.

Please note that tokens shown here are different from actual words in a sentence and will often contain compound words (समस्तपदानि) presented both as they appear as well as in a split manner.

For example, for the word “तपःस्वाध्यायनिरतम्” in Bala Kanda Sarga 1 Verse 1, you will see a total of FOUR tokens

1. तपःस्वाध्यायनिरतम् - Word as it appears in the corpus
2. तपस् - part of compound
3. स्वाध्याय - part of compound
4. निरतम् - part of compound

Anvaya

+

↺

?

तपस्वी

×

वाल्मीकिः

×

निरतम्

×

वरम्

×

नारदम्

×

पुंगवम्

×

तपःस्वाध्यायनिरतम्

×

तपस्

×

स्वाध्याय

×

वाग्निदा

×

वाच्

×

विदाम्

×

परिपप्रच्छ

×

मुनिपुंगवम्

×

मुनि

×

1

Submit

As shown above, the full word appears in blue, while the parts appear as yellow.

For such cases of compound words (समास), we need to retain only the full word, and unselect (by clicking on the x icon) the compound constituents.

i.e. After removing the partial tokens and rearranging, it would look as follows,

The screenshot shows the Anvaya interface. At the top, there's a header with the name 'Anvaya' and three icons: a red plus sign, a red circular arrow, and a grey question mark. Below the header, there's a list of tokens in a light green box. Each token is in a grey box with an 'x' icon to its right. The tokens are: तपस्वी, वाल्मीकिः, तपःस्वाध्यायनिरतं, वाग्विदां, वरम्, मुनिपुंगवम्, नारदम्, and परिपप्रच्छ. Below this list, there's a light red box containing the components of the selected token 'मुनिपुंगवम्'. The components are: तपस्, स्वाध्याय, निरतम्, पुंगवम्, वाच्, विदाम्, and मुनि. Each component is in a yellow box with a blue plus sign to its right. At the bottom left, there's a grey box with the number '1'. At the bottom right, there's a red 'Submit' button.

In most cases, sandhied words (सन्धियुक्तपदानि) are already presented as separate words, but this might not always be the case.

For example,

For the word “ममाप्येतत्”, from Ayodhya Kanda Sarga 18 Verse 2 (Verse ID 2489)

It is presented as two tokens “ममाप्य्” and “एतत्”. i.e. one of the sandhis is split, while the other is not. And, as before, for “ममाप्य्” the original token, as well as components, appear as separate tokens, i.e.

ममाप्य्	मम	अपि
---------	----	-----

Here, “ममाप्य्” is not a proper word, (i.e. made of two independent words, मम and अपि, unlike the previous example where component words were not standalone words).

Therefore, we want to keep the components instead. After annotation it should look like,

Anvaya

आर्ये × एतत् × मम × अपि × न × रोचते ×

ममाय् +

यत् × राघवः × राज्यश्रियं × त्यक्त्वा × वनम् × गच्छेत् ×

श्रियम् + राज्य +

2489 Submit

In case an Adhyaahara (अध्याहार) e.g. अस्ति is missing, such token can and should be added using the “+” button. This will present an interface as follows,

Add Token

**Mandatory**

Text: अस्ति

Lemma: अस्

**Analysis**

UPOS: UPOS

XPOS: XPOS

**Features**

Case: Case

Formation: Formation

Gender: Gender

Mood: Mood

Number: Number

Person: Person

Tense: Tense

VerbForm: VerbForm

Cancel Add

Anvaya

अस्ति ×

अस्मिन् × लोके × कः ×

Here, Text and Lemma are mandatory fields to be entered. After this, the token should appear on the right side in a dark colour. (e.g. अस्ति shown above). This can be dragged to the required sentence and arranged similar to other tokens.

It is important to note that even though words are shown as को or न् etc, the “Unsandhied” property gives the correct form of the word as “कः” and “न्” respectively and therefore these should not be removed.

## Named Entity Recognition (NER)

In this task, we identify tokens corresponding to named entities. We also assign categories to these entities.

### What are Named Entities?

Named entities are tokens (words) which are proper nouns of various things, such as humans, places, rivers, trees and so on.

e.g.

- राम - Name of a Human.
- अयोध्या - Name of a City.
- शरयू - Name of a River.
- रावण - Name of a Rakshasa.
- हनुमत् - Name of a Vaanara.

We have created a rich ontology with ~90 entity types. You have to select the most relevant entity type from these.

*NOTE 1: Entity labels refer to the “category” or “species” etc, and not to a “title”. i.e. we do not have “King” as an entity type because that is a title, and is temporal (namely, the same person, say Dasharatha, was “King” at some point and was no longer “King” at a later point).*

*NOTE 2: If you cannot find a suitable entity type, please reach out to us requesting a new type to be added.*

### How to annotate NER?

After Anvaya, you will be taken to NER task. On the right, a list of tokens will appear.

You can click on the (+) sign (button) in front of the tokens that qualify as named entities and select the type of entity from a dropdown in front of it.



e.g. In Ayodhya Kanda, Sarga 18, Verse 1,





तथा तु विलपन्ती तां कौसल्यां राममातरम्  
उवाच लक्ष्मणो दीनस् तत् कालसदृशं वचः


There are two named entities, कौसल्या (appearing as कौसल्याम्) and लक्ष्मण (appearing as लक्ष्मणो in the verse and लक्ष्मणः in the token list) as shown below. The category for these has been selected as Human.

Named Entity Recognition




लक्ष्मणः	HUMAN Human	
कौसल्याम्	HUMAN Human	


तथा




तु




दीनः



ताम्



विलपतीम्



2488

Submit

## Coreference Resolution

### What are coreferences?

Coreferences are pairs of words (tokens) that refer to the same physical entity (coreferences). This typically manifests in the form of pronouns.

Consider the first two verses from Ayodhya Kanda Sarga 18.

तथा तु विलपन्तीम् तां कौसल्यां राममातरम् ।  
उवाच लक्ष्मणो दीनस् तत् कालसदृशं वचः ॥  
न रोचते मम अपि एतद् आर्ये यद् राघवो वनम् ।  
त्यक्त्वा राज्यश्रियं गच्छेत् स्त्रिया वाक्यवशं गतः ॥

Here, the words styled in a similar manner refer to the same physical entity.

i.e. Those in red all refer to Kausalya.

**It is important to understand that by coreference we mean two or more expressions that refer to the same person or thing. It does NOT mean all words in the same vibhakti.**

For example, कौसल्यानंदन, दशरथात्मज, जानकीवल्लभ etc. refer to राम and hence should be marked as coreferents. But वीरः, मुदितः, खिन्नः etc. are generic adjectives and cannot be co-referents.

In the first verse of Balakanda -- “तपःस्वाध्यायनिरतं तपस्वी नारदं मुनिपुंगवम्”

All the underlined words are in the same vibhakti (Dwitiya) but तपःस्वाध्यायनिरतं or मुनिपुङ्गवम् **should not** be marked as a coreferent of नारद.

### How to annotate?



In this task, you will be presented with a list of tokens as clickable buttons. More often than not, a pronoun is used in subsequent sentences after the mention of a noun.

Therefore, in addition to the current verse, you will also be shown sentences from upto five previous verses.

Once you identify a co-referring pair, say, ताम् and कौसल्याम् you can click on the tokens one after the other and they will start appearing in the yellow row. Click on the green check mark to confirm a pair. This will empty the yellow row and move the pair in the confirmed list of coreferences. You can add more coreference pairs like this. After adding all the coreference pairs, click on “Submit”.

NOTE: While annotating a verse, coreference pairs from previous verses will also be visible to you for reference and to avoid repetition.

Co-reference Resolution



तथा तु दीनः लक्ष्मणः ताम् विलपतीम् राममातरम् कौसल्याम् तत् कालसदृशं

वचः उवाच

आर्ये एतत् मम अपि न रोचते

यत् राघवः राज्यश्रियं त्यक्त्वा वनम् गच्छेत्

↔

↻

✓

मम

↔

लक्ष्मणः

—

आर्ये

↔

कौसल्याम्

—

राममातरम्

↔

कौसल्याम्

—

ताम्

↔

कौसल्याम्

—

2489

Submit

## Contact Us

If you face any issues, you may submit them using the Google form:

<https://forms.gle/sKSE7M5oGXneicEf9>

Additionally, in case of any difficulty, you can reach out to us using e-mail. (Remove spaces from e-mail IDs)

- Hrishikesh Terdalkar: [hrishirt@cse.iitk.ac.in](mailto:hrishirt@cse.iitk.ac.in)
- Chaitali Dangarikar: [cadangarikar@gmail.com](mailto:cadangarikar@gmail.com)