

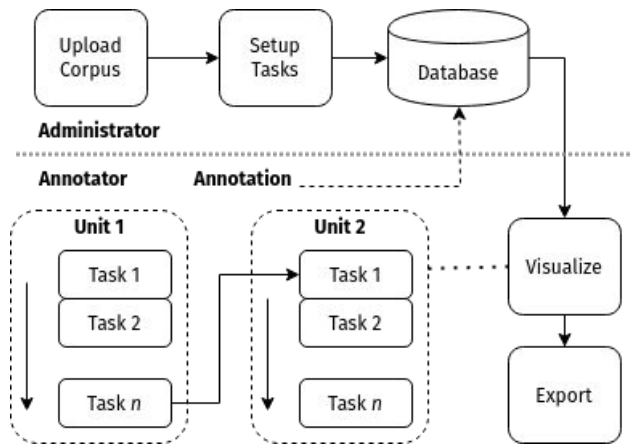
# Antarlekhaka: A Comprehensive Tool for Multi-task Natural Language Annotation

~ 7000 Low-resource Language  $\Rightarrow$  Human annotation relevant!

- Ambiguous/Absent sentence boundaries
- Ordering, splitting, merging tokens
- Limited support in existing tools

- Corpora in poetry format
- Majority of Sanskrit literature in poetry

The screenshot shows the Antarlekhaka interface. At the top, there is a 'Sentence Boundary' section with a text input field containing '2488 तथा तु विलपन्तीं तां कौसल्यां राममातरम् उवाच लक्ष्मणो दीनस् तत् कालसदृशं वचः ##'. Below this, there is a 'Canonical Word Order' section with a text input field containing '24 ममाप्यु' and a list of words: 'आर्ये', 'एतत्', 'मम', 'अपि', 'न', 'रोचते'. A 'Submit' button is at the bottom right of the interface.



## Eight Categories of NLP Tasks

- Sentence Boundary**
- Token Manipulation**
- Token Text Annotation
- Token Classification
- Token Graph
- Token Connection
- Sentence Classification
- Sentence Graph

- Sequential + Multi-task
- Pluggable heuristics
- Language agnostic
- Clone, Visualize, Export
- Administrative tasks



Hrishikesh Terdalkar  
Arnab Bhattacharya



[Antarlekhaka/code](https://github.com/Antarlekhaka/code)  
[hrishikeshrt.github.io](https://github.com/hrishikeshrt.github.io)

