VPCT Annotation Guidelines

Hello, dear annotator, and thank you for your time!
Here is a brief background for the task and the specific annotation guidelines. This task is a part of the DharmaBench paper and evaluation benchmark for Sanskrit and Tibetan.

Introduction

Task: Classifying verse from prose.

Goal: passage-level classification: categorizing a text chunk as predominantly verse-based or prose-based compositions.

Background: This task requires passage-level classification to distinguish predominantly verse-based from prose-based passages, which helps separate cited root-text verses from prose commentaries. Annotators should rely on formal literary features—such as lineation, syllable patterns, and terminal markers—rather than semantic content when labeling samples.

Real-world relevancy: Specifically, this task can help distinguish between cited “root-text” passages when these are composed in verse and commentary sections when these are written in prose, based on formal literary features rather than content analysis—the latter being Task 6’s ultimate objective (which enables the exclusion of verses that are not from the “root text”).

The main formal features of Tibetan verse include several distinctive characteristics. Lines typically contain an odd number of syllables, with seven being the most common, while lines exceeding fifteen syllables are rare. Grammatical particles tend to occupy even-numbered positions (2nd, 4th, 6th, and so forth), creating a rhythmic pattern throughout the verse. Each line concludes with a double stroke that serves as a terminal marker.

Example

1. Text: "།མ་འདར་ཨ་ཉ་དག་ན་ར་རྣ་ལ་ཞ་རན་ཅམ་ཡང་འཇ་གི་ར་ཀྱ་ན་དགི་ན་ཐུ་ན་མ་ངམ་ན། ། ལ་གས་གནས་འཁ་ལ་འ་སཚན་པ་དང། ། ཡངས་ཤ་ད་ཞབས་ལ་ང་མ་མང་ན་དང། །མ་ར་མ་ར་ང་དང་ཞ་ཐབས་ན་།" - Class: Verse
2. Text: "།ཡ་ཤ་མ་དབང་ཐ་བ་གཡུ་ར་པ་ད་དགལ། ། བམ་ད་ན་མས་འད་ཁ་པའ་གང་ཟག་ལ་ལ་ཞ་གི། བམཀ་ལ་པ་བཡ་བ་ཁ་གཏ་ཁ་ག་བར་མཆ་ད་བཡ་ད། །" - Class: Prose

Reichman University Data Science Institute

INTELLEXUS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Data collection guidelines

1. Go over texts from the defined sources list (Works written in verse and prose from the ACIP and rKTs Derge Kangyur and Tengyur, along with autochthonous works from online sites like Tsadra).
2. Collect the texts written (predominantly) in verse and prose and separate them.
3. Sub-classify the verse works into two folders: autochthonous materials and allochthonous materials. Do the same for the prose work.
4. Share the folder containing these materials with the responsible RUNI data scientist.

Additional note (folder-based workflow):
We will use pre-structured folders for annotation. Each folder corresponds to a defined class (VERSE or PROSE).
Annotators should review their assigned folders, remove non-relevant or corrupted files, and ensure only valid text samples remain before beginning annotation.

Annotation guidelines

Annotation platform

We will be using the Label Studio app. We will upload the collected CSV files and set up the annotation project.

Preparation

1. Log in to our Label Studio organization here.
2. Let us know, and we will approve your user and assign you to the respective project.

Annotation guidelines

1. Read the sentence/paragraph thoroughly.
2. Ensure the label remains intact - it is pre-annotated with the source's metadata.
3. Annotation Speed vs. Accuracy Tradeoff: Emphasize **accuracy over speed** if necessary.

Something is Off?

Reach out to us!
If you encounter technical problems with the folder setup, missing texts, or conceptual ambiguities, please contact the project leads.

Thank you very much for your contribution!