AACT 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: Allochthonous vs. Autochthonous Classification Tibetan (AACT)

Goal: To distinguish between Tibetan allochthonous works and autochthonous works; ultimately, to determine the origin of a given work with unknown or uncertain provenance.

Background: This downstream task classifies Tibetan texts into two categories: allochthonous works (ALLO), which are texts translated primarily from Sanskrit, and autochthonous works (AUTO), which are indigenous Tibetan compositions. The task is challenging for three main reasons: 1) Autochthonous works frequently borrow, cite, or adapt material from allochthonous sources; 2) The genres and topics of both classes are intricate and often overlapping; texts within the same or similar genre or topic but belonging to different classes may exhibit greater similarity than texts from the same class but different genres/topics (e.g., a commentary in ALLO may more closely resemble a commentary in AUTO than it would a prayer in ALLO); 3) In the context of Tibetan Buddhist literature, Indic origin is generally regarded as conferring greater authority and authenticity; consequently, some AUTO works intentionally imitate the style of ALLO, and some disputed cases claim allochthonous provenance.

Real-world relevancy: This task is particularly significant as it will ultimately support the identification of linguistic differences between allochthonous and autochthonous Classical Tibetan, including their distinctive stylistic features in grammar, syntax, and vocabulary. In doing so, it could serve as a foundation for other tasks, particularly those concerning authorship analysis, including attribution, verification, distinction, clustering, and author profiling. It potentially could also serve as a foundation for translation analysis, including classification of the allochthonous texts into “Ancient translations” (mainly mid 10th – mid 9th cent) and “New translations” (mainly late 10th – 13th cent), and even translator’s profiling.

Example

1. Text: "པའི་དོན་གསུངས་སོ། | ལེའུའི་མཚན་གྱི་མཇུག་བསྡུ་བ་ནི་དེ་བཞིན་གཤེགས་པ་ཐམས་ཅད་ཀྱི་་སྙིང་པོ་བསྐུལ་བ་ཞེས་པ་དོན་དངོས་གྲུབ་ཀྱི་དོན་མ་ཐོབ་པ་ཐོབ་པར་བྱེད་ཅིང་། ཐོབ་པའི་ནུས་པ་བརྟན་ཞིང་ཕྱིར་མི་ལྡོག་པར་གྱུར་པས་ན་གནས་སྐབས་དང་་མཐར་ཐུག་གི་འབྲས་བུ་ཁྱད་པར་ཅན་ཐོབ་པའི་དོན་ཏོ། །ལེའུ་བཅུ་པའི་རྣམ་པར་་བཤད་པའ" - Class: AUTO
2. Text: "༄༅། །དཔའི་བསྔོ་བ་དེ་དུག་མེད་པ་ཡིན་ནོ། །རིགས་ཀྱི་བུའམ། རིགས་ཀྱི་བུ་མོ་ཐེག་པ་ཆེན་པོ་ལ་ཡང་དག་པར་ཞུགས་པ་གང་མཚན་མའི་ཚུལ་དང་། དམིགས་པའི་ཚུལ་གྱིས་དགེ་བའི་རྩ་བ་དེ་དག་ཡོངས་སུ་བསྔོ་བར་བྱེད་པ་དེ་ནི། དགེ་བའི་རྩ་བ་དེ་དག་ལོག་པར་ཡོངས་སུ་བསྔོ་བ་ཡིན་ཏེ། ཡང་དག་པར་ཡོངས་སུ་བསྔོ་བ་མ་ཡིན་ནོ། | ལོག་པར་ཡོངས་སུ་བསྔོ་བ་ལ་ནི། སངས་རྒྱས་བཅོམ་ལྡན་འདས་རྣམས་བསྔགས་པ་མི་མཛད་དོ། །གང་སངས་རྒྱས་བཅོམ་ལྡན་འདས་རྣམས་ཀྱིས་བསྔགས་པ་མ་མཛད་པའི་ཡོངས་སུ་བསྔོ་བས་ཡོངས་སུ་བསྔོ་བ་དེ་ནི། སྦྱིན་པའི་ཕ་རོལ་ཏུ་ཕྱིན་པ་ཡོངས་སུ་རྫོགས་པར་མི་བྱེད་དོ། །ཚུལ་ཁྲིམས་ཀྱི་ཕ་རོལ་ཏུ་ཕྱིན་པ་དང་། བཟོད་པའི་ཕ་རོལ་ཏུ་ཕྱིན་པ་དང་། བརྩོན་འགྲུས་ཀྱི་ཕ་རོལ་ཏུ་ཕྱིན་པ་དང་། བསམ་གཏན་གྱི་ཕ་རོལ་ཏུ་ཕྱིན་པ་དང་། ཤེས་རབ་ཀྱི་ཕ་རོལ་ཏུ་ཕྱིན་པ་ཡོངས་སུ་རྫོགས་པར་མི་བྱེད་དོ། །གང་ཕ་རོལ་ཏུ་ཕྱིན་པ་དྲུག་པོ་དག་ཡོངས་སུ་རྫོགས་པར་མི་བྱེད་དོ། | བྱང་ཆུབ་ཀྱི་ཕྱོགས་ཀྱི་ཆོས་སུམ་ཅུ་རྩ་བདུན་ཡོངས་སུ་རྫོགས་པར་མི་བྱེད་དོ། །ནང་སྟོང་པ་ཉིད་དང་། དངོས་པོ་མེད་པའི་ངོ་བོ་ཉིད་སྟོང་པ་ཉིད་ཀྱི་བར་དང་། སྟོབས་རྣམས་དང་། མི་འཇིགས་པ་རྣམས་དང་། སོ་སོ་ཡང་དག་པར་རིག་པ་རྣམས་དང་། སངས་རྒྱས་ཀྱི་ཆོས་མ་འདྲེས་པ་བཅོ་བརྒྱད་ཡོངས་སུ་རྫོགས་པར་མི་བྱེད་དོ། །དེ་ནི། སངས་རྒྱས་ཀྱི་ཞིང་ཡོངས་སུ་དག་པར་མི" - Class: ALLO

Data Collection Guidelines

1. For ALLO, first download the entire cannon (sources: ACIP and Esukhia Derge Kangyur for scriptural works; ACIP Derge Tengyur for non-scriptural works).
2. Apply a labeling mechanism, including necessary sub-classes.
3. Retain only those works for which there is evidence of Sanskrit manuscripts, fragments, or references in other Indic sources. Translations from Chinese should also be excluded.
4. For AUTO, select works with non-controversial authorship from both the early and later periods of the Dissemination of Dharma in Tibet, covering a wide range of genres and topics.
5. Apply a labeling system. One may consult the ALLO label from Step 2 as needed.
6. The same structured datasets can be used for tasks SCCT and THCT.

Additional note (folder-based workflow):

We will use pre-structured folders for annotation. Each folder corresponds to a defined class (ALLO or AUTO).

Annotators should review their assigned folders, remove non-relevant or corrupted files, and ensure only valid text samples remain before beginning annotation.

Annotation Guidelines

1. Read the sample text carefully.
2. Reject the sample if it contains substantial material that is not part of the original text, such as tables of contents, editorial notes, translator’s remarks and so forth.
3. Delete minor extraneous material not part of the original text, such as Chinese characters or other trivial elements.
4. Reject the sample if it is too short (i.e., fewer than two complete sentences).
5. For conversion errors (i.e. due to incorrect input in the original material), correct minor ones. Reject the sample if there are numerous errors that cannot be easily resolved.
6. In cases of ambiguity, consult with the team and your supervisors. If an agreement cannot be reached, discard the sample.

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!