# QUDT 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:** Identify explicit citations and quotes, along with the title of the source and/or the name of the author (if available), in Tibetan (QUDT)

**Goal:** Correct detection of citations (and the parts of the citations, as per Entities below) in the text

**Background:** Tibetan literature extensively employs quotes and citations. This can, for example, reinforce or criticize an opinion or argument, or draw attention to specific topics and themes.

1. This task is curated under the Intellexus project. Its goal is to assess LLMs' performance on this task, namely identifying explicit citations, along with the title of the source and/or the name of the author (if available).
2. It can be challenging because citations can appear in a variety of formats, and sometimes texts may contain spelling mistakes/typos or lack information such as the source or author being cited.

**Real-world relevancy**: Identifying citations across a large corpus of Tibetan texts represents the foundational step toward understanding the intellectual lineage, textual transmission, and scholarly authority in Tibetan literature. This analysis could facilitate future research endeavors, such as tracing the intellectual lineages of specific ideas and the intertextual dialogue between different philosophical schools, as evidenced by their citation practices. When combined with other analytical and computational capabilities, this approach can also help map the historical development of key Buddhist concepts, including the authoritative sources used to substantiate them and the transformation of these canonical references over time.

# Annotation process

## Basic instructions

1. Log in to (annotation platform)
2. Read the text thoroughly...
3. Identify the dominant sentiment...
4. Assign a label from the below-listed list
5. Re-read and re-evaluate your annotation.
6. Annotation Speed vs. Accuracy Tradeoff: Emphasize accuracy over speed if necessary

## Annotation Phase 1 - Exploratory Annotation

After initially defining the annotation guidelines with cooperation from Tibetan team, we presume that due to the noisy nature of data there will be “edge cases” which the guidelines do not cover.

Therefore for the first 80 samples, each of the annotators will begin by labelling all 80 samples separately (so 80*N annotations if we have N annotators). For samples with less than 95% agreement, the annotators will decide together the “correct” annotation, discuss why the guidelines didn’t cover this case (why we had diverging annotations), and update the guidelines to cover this case.

## Annotation Phase 2 - Exploratory Annotation II

In this phase, several new annotators joined the team. Thus, for an additional 40 samples, we apply the same process as in phase 1 on the *updated* guidelines (in total, about 10% of all data). This phase served to both validate the improvement in annotator agreement by using the updated guidelines, and also to calibrate the new annotators on how to annotate correctly.

## Annotation Phase 3 - Round Robin Annotation

The remaining samples are distributed randomly, and each person takes “shifts” between annotating and reviewing other people’s annotations. For example, in the first half of the week, person X annotates and person Y reviews; in the second half, they switch. This ensures that every annotator is familiar with every other annotator’s annotation style, while also addressing real-world logistical constraints, such as limited numbers of annotators and large volumes of data.

If an annotator and a reviewer disagree, they discuss the sample with the entire group, reach a conclusion, and update the guidelines accordingly.

In addition, the group regularly convenes to discuss annotation, ensuring a consistent annotation style among annotators.

# WORKING ANNOTATION GUIDELINES

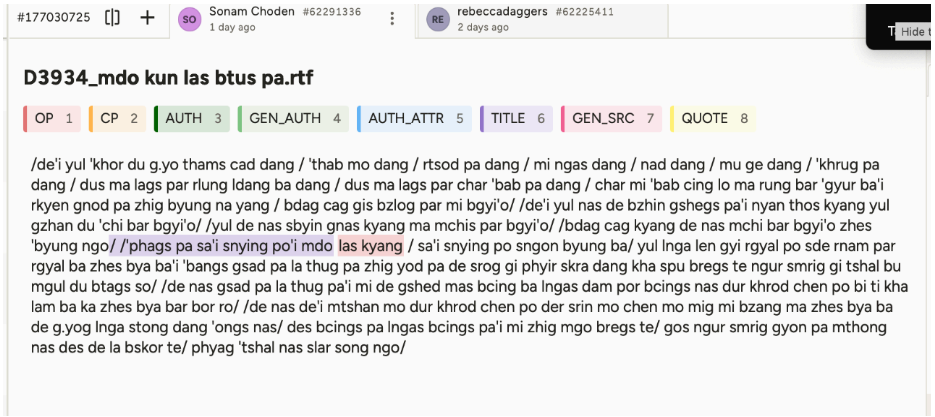
Continuously update the guidelines below as needed, using the agreed-upon methodology.

## Labels

| **Entity** | **Short Description** | **Examples** |
|---|---|---|
| OP (opening phrase/particle) | Opening Particle marking the start of a citation | las/nas/la don (gsungs pa)<br><br>ji skad du<br><br>na re<br><br>las shes te<br>la ’byung ba<br>las kyang<br>de ltar yang<br><br>ergative (after AUTH / GEN_AUTH)<br>pas bka’ stsal pa |
| CP (closing phrase/particle) | Closing Particle marking the end of a citation | ces/shes/zhes (bya ba / verb such as gsungs + pa/ba + particle, in case of gsungs pa lta bu’o/bzhin no, do include the lta bu’o, and if there is a longer closing phrase, for example “zhes ji skad du gsungs pa bzhin no/” then mark the whole phrase)<br>zhes gsungs so<br>zhe'o<br>zhe na (only when it is clearly a citation and not just an objection or similar) |
| AUTH (author) | The name of the author being cited | if it is *bcom ldan ’das* without a cited source, mark it as author. If a source is provided, it can be ignored.. eg. *yang bcom ldan 'das kyis lang kar gshegs pa de nyid las bka' stsal pa/* |

| | | |
|---|---|---|
| GEN_AUTH (generic author) | The name of the author is not provided, but only a generic indication of the author or speaker, with words like “some [say]”, “others [say]”, “my teacher [says]”, etc. | kha cig / gzhan<br>de skad?<br>what if it is “I say”? |
| AUTH_ATTR (author attribute) | Pre- and post-positive attributes | slob dpon / paṇḍi ta / zhabs / dpal ldan, khad par dran par dang bral ba, sla na med pa’i slon dpon, etc. |
| TITLE (title) | The title of the text being cited | rig pa mchog gi rgyud chen po<br>nam mkha’i snying po’i le’u |
| GEN_SRC (generic source) | No title/author of the text being cited is provided, but only a generic indication of the source, with words like “[in] the same [text]”, “[in] another [text]”, “[in] the sūtra”, etc. | ’di nyid / gzhan / mdo / rgyud,  etc. |
| QUOTE (citation text) | The text being cited | sku gzugs la thang dkar blta zhing bsgrub par bya ste\| gzhan du lta zhing bsgrub par mi bya’o |
| INVALID | Please contact the RUNI team if you think a text is invalid | |

**Important Note: We agreed to NOT ignore the shad in general.**



D3934_mdo kun las btus pa.rtf

/de'i yul 'khor du g.yo thams cad dang / 'thab mo dang / rtsod pa dang / mi ngas dang / nad dang / mu ge dang / 'khrug pa dang / dus ma lags par rlung ldang ba dang / dus ma lags par char 'bab pa dang / char mi 'bab cing lo ma rung bar 'gyur ba'i rkyen gnod pa zhig byung na yang / bdag cag gis bzlog par mi bgyi'o/ /de'i yul nas de bzhin gshegs pa'i nyan thos kyang yul gzhan du 'chi bar bgyi'o/ /yul de nas sbyin gnas kyang ma mchis par bgyi'o/ /bdag cag kyang de nas mchi bar bgyi'o zhes 'byung ngo/ /'phags pa sa'i snying po'i mdo las kyang / sa'i snying po sngon byung ba/ yul lnga len gyi rgyal po sde rnam par rgyal ba zhes bya ba'i 'bangs gsad pa la thug pa zhig yod pa de srog gi phyir skra dang kha spu bregs te ngur smrig gi tshal bu mgul du btags so/ /de nas gsad pa la thug pa'i mi de gshed mas bcing ba lngas dam por bcings nas dur khrod chen po bi ti kha lam ba ka zhes bya bar bor ro/ /de nas de'i mtshan mo dur khrod chen po der srin mo chen mo mig mi bzang ma zhes bya ba de g.yog lnga stong dang 'ongs nas/ des bcings pa lngas bcings pa'i mi zhig mgo bregs te/ gos ngur smrig gyon pa mthong nas des de la bskor te/ phyag 'tshal nas slar song ngo/

# Edge cases and common mistakes

1. Clearly define what to do in edge cases.
2. Annotate INVALID if a sample

   is incomplete.

   Doesn't meet the quality criteria.

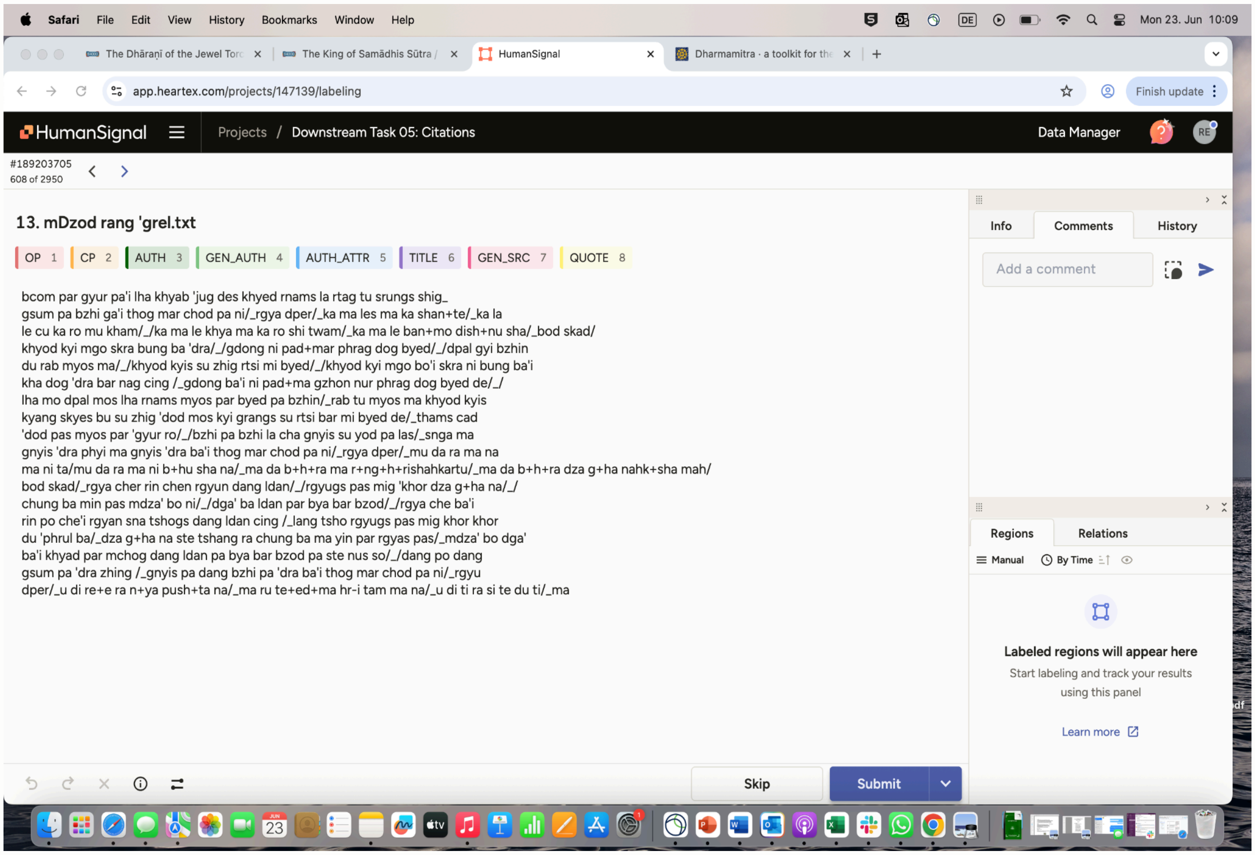
   Contains an unknown transliteration or a different language than the intended one.
3. In the case of ambiguity, please consult the team.
4. If a sample can be classified as more than one label, ping the team and discuss how to treat this case. Update the guidelines accordingly to make them no longer ambiguous, and add the example to the Examples section.

Opening particles - OP can vary, for example, sometimes after las there can be gsungs pa

In some cases, we might encounter OP such as the following:

ogs pa gsungs pa bzhin no/ /'**phags pa blo gros mi zad pa'i mdo las kyang /** sdug bsngal gyi tshor bas reg na ngan song dang mi khom par skyes pa'i sems can thams cad la snying rje chen po bskyed do/ /**de bzhin du sbyar te/ tshor ba ni mngon par zhen pa'o/ /tshor ba ni yongs su 'dzin pa'o/ /tshor ba ni nye bar len pa'o/ /tshor ba ni dmigs pa'o/ /tshor ba ni phyin ci log go/ /tshor ba ni rnam par rtog pa'o zhes bya ba la sogs pa gsungs so/ /**chos yang dag par sdud pa'i mdo las kyang / tshor ba myong bar rab bstan pa/ /tshor ba las gzhan tshor ba po/ /gud na yod pa ma yin na/

b) See the screenshot below: it appears that there are quotes, but they are introduced as “examples”, as in “rgya dper…” - they do appear to be proper quotes, but I am not sure if we want to include these sorts of examples?


Safari File Edit View History Bookmarks Window Help
Mon 23. Jun 10:09
The Dhāraṇī of the Jewel Torc
The King of Samādhis Sūtra /
HumanSignal
Dharmamitra · a toolkit for the
app.heartex.com/projects/147139/labeling
Finish update
HumanSignal
Projects / Downstream Task 05: Citations
Data Manager
#189203705
608 of 2950
Info
Comments
History
Add a comment
Regions
Relations
Manual
By Time
Labeled regions will appear here
Start labeling and track your results using this panel
Learn more
Skip
Submit

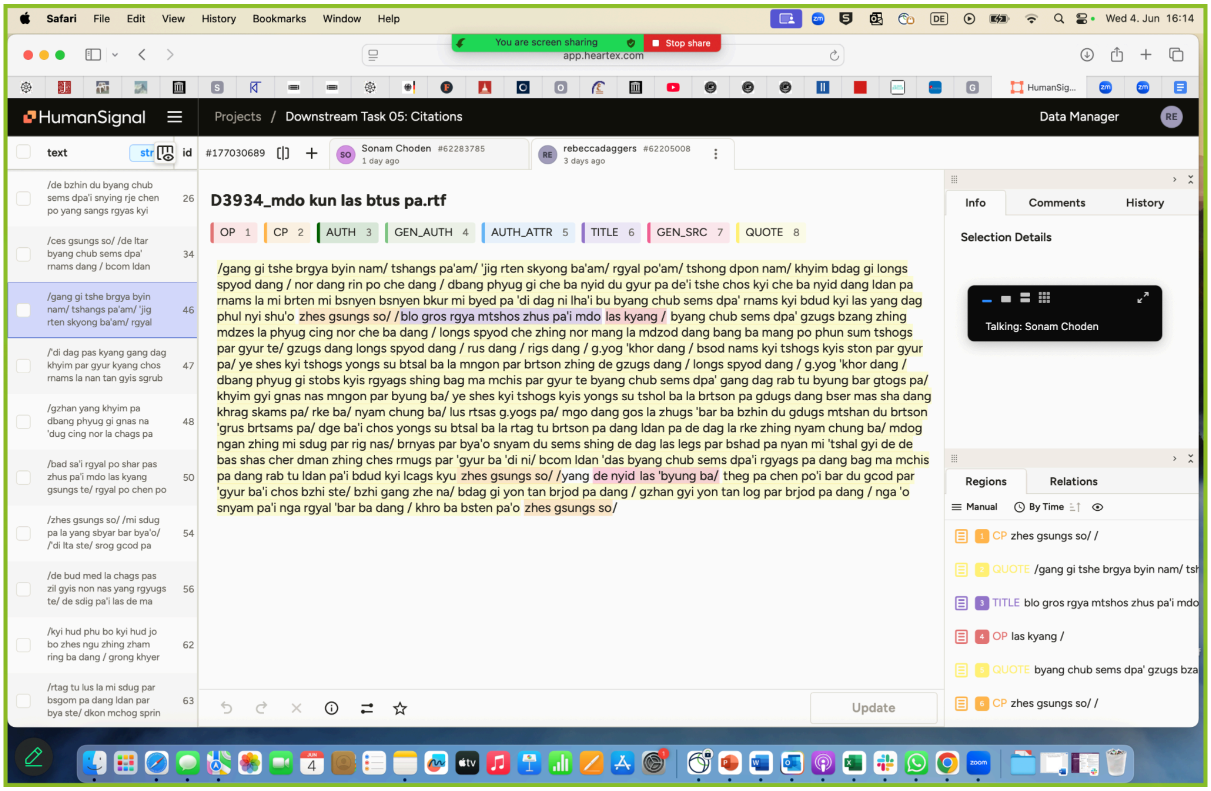
13. mDzod rang 'grel.txt

| OP 1 | CP 2 | AUTH 3 | GEN_AUTH 4 | AUTH_ATTR 5 | TITLE 6 | GEN_SRC 7 | QUOTE 8 |
|---|---|---|---|---|---|---|---|

bcom par gyur pa'i lha khyab 'jug des khyed rnams la rtag tu srungs shig_
gsum pa bzhi ga'i thog mar chod pa ni/_rgya dper/_ka ma les ma ka shan+te/_ka la
le cu ka ro mu kham/_/ka ma le khya ma ka ro shi twam/_ka ma le ban+mo dish+nu sha/_bod skad/
khyod kyi mgo skra bung ba 'dra/_/gdong ni pad+mar phrag dog byed/_/dpal gyi bzhin
du rab myos ma/_/khyod kyis su zhig rtsi mi byed/_/khyod kyi mgo bo'i skra ni bung ba'i
kha dog 'dra bar nag cing /_gdong ba'i ni pad+ma gzhon nur phrag dog byed de/_/
lha mo dpal mos lha rnams myos par byed pa bzhin/_rab tu myos ma khyod kyis
kyang skyes bu su zhig 'dod mos kyi grangs su rtsi bar mi byed de/_thams cad
'dod pas myos par 'gyur ro/_/bzhi pa bzhi la cha gnyis su yod pa las/_snga ma
gnyis 'dra phyi ma gnyis 'dra ba'i thog mar chod pa ni/_rgya dper/_mu da ra ma na
ma ni ta/mu da ra ma ni b+hu sha na/_ma da b+h+ra ma r+ng+h+rishahkartu/_ma da b+h+ra dza g+ha nahk+sha mah/
bod skad/_rgya cher rin chen rgyun dang ldan/_/rgyugs pas mig 'khor dza g+ha na/_/
chung ba min pas mdza' bo ni/_/dga' ba ldan par bya bar bzod/_/rgya che ba'i
rin po che'i rgyan sna tshogs dang ldan cing /_lang tsho rgyugs pas mig khor khor
du 'phrul ba/_dza g+ha na ste tshang ra chung ba ma yin par rgyas pas/_mdza' bo dga'
ba'i khyad par mchog dang ldan pa bya bar bzod pa ste nus so/_/dang po dang
gsum pa 'dra zhing /_gnyis pa dang bzhi pa 'dra ba'i thog mar chod pa ni/_rgyu
dper/_u di re+e ra n+ya push+ta na/_ma ru te+ed+ma hr-i tam ma na/_u di ti ra si te du ti/_ma

# Examples

Attach a fully annotated example. Include edge cases.



| text | id |
|---|---|
| /de bzhin du byang chub sems dpa'i snying rje chen po yang sangs rgyas kyi | 26 |
| /ces gsungs so/ /de ltar byang chub sems dpa' rnams dang / bcom ldan | 34 |
| /gang gi tshe brgya byin nam/ tshangs pa'am/ 'jig rten skyong ba'am/ rgyal | 46 |
| /'di dag pas kyang gang dag khyim par gyur kyang chos rnams la nan tan gyis sgrub | 47 |
| /gzhan yang khyim pa dbang phyug gi gnas na 'dug cing nor la chags pa | 48 |
| /bad sa'i rgyal po shar pas zhus pa'i mdo las kyang gsungs te/ rgyal po chen po | 50 |
| /zhes gsungs so/ /mi sdug pa la yang sbyar bar bya'o/ /'di lta ste/ srog gcod pa | 54 |
| /de bud med la chags pas zil gyis non nas yang rgyugs te/ de sdig pa'i las de ma | 56 |
| /kyi hud phu bo kyi hud jo bo zhes ngu zhing zham ring ba dang / grong khyer | 62 |
| /rtag tu lus la mi sdug par bsgom pa dang ldan par bya ste/ dkon mchog sprin | 63 |

#177030689

**D3934_mdo kun las btus pa.rtf**

OP 1 | CP 2 | AUTH 3 | GEN_AUTH 4 | AUTH_ATTR 5 | TITLE 6 | GEN_SRC 7 | QUOTE 8

/gang gi tshe brgya byin nam/ tshangs pa'am/ 'jig rten skyong ba'am/ rgyal po'am/ tshong dpon nam/ khyim bdag gi longs spyod dang / nor dang rin po che dang / dbang phyug gi che ba nyid du gyur pa de'i tshe chos kyi che ba nyid dang ldan pa rnams la mi brten mi bsnyen bsnyen bkur mi byed pa 'di dag ni lha'i bu byang chub sems dpa' rnams kyi bdud kyi las yang dag phul nyi shu'o zhes gsungs so/ /blo gros rgya mtshos zhus pa'i mdo las kyang / byang chub sems dpa' gzugs bzang zhing mdzes la phyug cing nor che ba dang / longs spyod che zhing nor mang la mdzod dang bang ba mang po phun sum tshogs par gyur te/ gzugs dang longs spyod dang / rus dang / rigs dang / g.yog 'khor dang / bsod nams kyi tshogs kyis ston par gyur pa/ ye shes kyi tshogs yongs su btsal ba la mngon par brtson zhing de gzugs dang / longs spyod dang / g.yog 'khor dang / dbang phyug gi stobs kyis rgyags shing bag ma mchis par gyur te byang chub sems dpa' gang dag rab tu byung bar gtogs pa/ khyim gyi gnas nas mngon par byung ba/ ye shes kyi tshogs kyis yongs su tshol ba la brtson pa gdugs dang bser mas sha dang khrag skams pa/ rke ba/ nyam chung ba/ lus rtsas g.yogs pa/ mgo dang gos la zhugs 'bar ba bzhin du gdugs mtshan du brtson 'grus brtsams pa/ dge ba'i chos yongs su btsal ba la rtag tu brtson pa dang ldan pa de dag la rke zhing nyam chung ba/ mdog ngan zhing mi sdug par rig nas/ brnyas par bya'o snyam du sems shing de dag las legs par bshad pa nyan mi 'tshal gyi de de bas shas cher dman zhing ches rmugs par 'gyur ba 'di ni/ bcom ldan 'das byang chub sems dpa'i rgyags pa dang bag ma mchis pa dang rab tu ldan pa'i bdud kyi lcags kyu zhes gsungs so/ /yang de nyid las 'byung ba/ theg pa chen po'i bar du gcod par 'gyur ba'i chos bzhi ste/ bzhi gang zhe na/ bdag gi yon tan brjod pa dang / gzhan gyi yon tan log par brjod pa dang / nga 'o snyam pa'i nga rgyal 'bar ba dang / khro ba bsten pa'o zhes gsungs so/

1 CP zhes gsungs so/ /
2 QUOTE /gang gi tshe brgya byin nam/ tsh
3 TITLE blo gros rgya mtshos zhus pa'i mdo
4 OP las kyang /
5 QUOTE byang chub sems dpa' gzugs bza
6 CP zhes gsungs so/ /

Note that the partial quotes are marked as long as there is a closing or opening particle.

# Q & A

1. How and where to report problematic cases, unclear guidelines, or dataset inconsistencies?
    a. Ping the team and discuss together how to treat this case. Once a unanimous conclusion and agreement has been reached, update the guidelines accordingly, **and add this example to the Examples section** above.
2. What to do if the annotation platform has issues?
    a. The UHH team should ping Ari on Slack. If the issue is urgent and requires response within 24 hours, message him via Whatsapp.

# Reviewing Guidelines

Reviewing is very different from annotating, because you look for slightly different things.

## Reviewing Goals:

When you are **annotating**, your main goal is to detect the different sections of each sample (e.g., tags, such as Author, Quote, OP, CP, etc.) in the Citations Task.

When you are **reviewing**, your main goal is to challenge the existing detection and ensure it meets the guidelines. This means that you want to answer the following questions:

**A)** Does this annotation perfectly align with the guidelines we defined in the annotation document?

**B)** If there are multiple annotations for this sample, which one of these is "more correct"? Ideally, the reason why one annotation is better than another shouldn't be up to intuition; it should be based on which one follows the guidelines better. If there is a 'gray zone' that is not explicitly defined, please discuss it, update the guidelines with examples, and notify the team.

**B.2)** If there are multiple annotations for this sample, select one of these samples to be the ground truth.

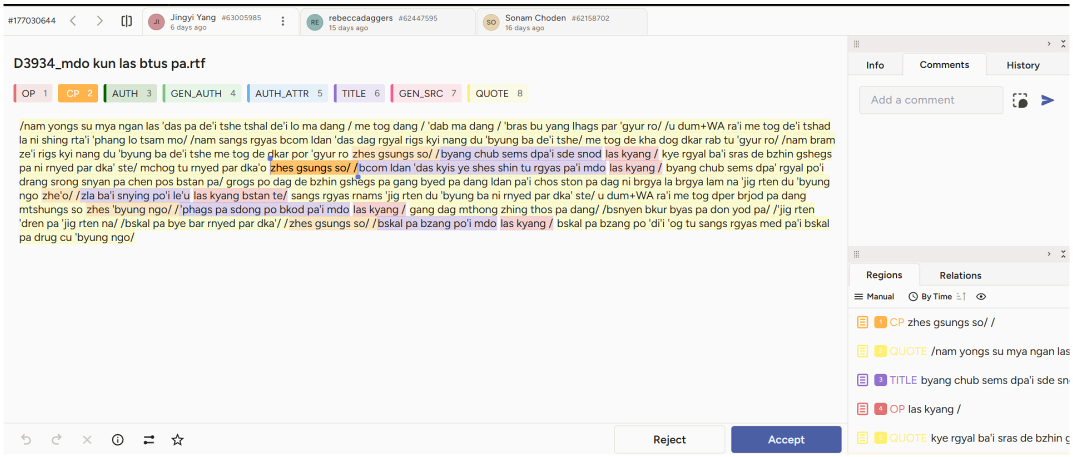
**C) Important:** Does this annotation contain labelling mistakes? This can include two different labels overlapping each other by accident, a part of a word accidentally not labelled when it should be, a label extending to the neighboring word when it shouldn't, et cetera. These can be hard to spot if you are not deliberately scanning for them!

## Reviewing In LabelStudio: How-To

To review, start by clicking on the Review button from the Project tab. This will show you existing samples that have at least one annotation. If multiple people have annotated it, you can navigate to their annotations in the top; they have a tab (see picture).

- An annotation can either be accepted or rejected. The goal is to have a single ground truth for each sample.
- If an annotation has a tiny mistake that doesn't take much effort to fix, you can update it by clicking on the existing label and dragging one of its edges to make it longer/shorter, much like when annotating.
- If an annotation has many mistakes and takes more than a minute or two to fix, you can leave a comment and reject the annotation, and this will appear in the data manager. Please notify the annotator if you reject their annotation, and send them the sample ID.
- If there are multiple valid annotations, please select one of them as the ground truth by clicking on the three dots and selecting "Select ground truth". Otherwise, it will assume an annotation is valid if you reject all other annotations for this sample.
- If there is a single annotation, it will assume it is valid once you have accepted it.

**Unlike annotation**, there is nobody to check your work when you are a reviewer, so it is **extremely important** to be careful and alert - always challenge the sample you are seeing, even if it seems obvious or if it's an empty sample.

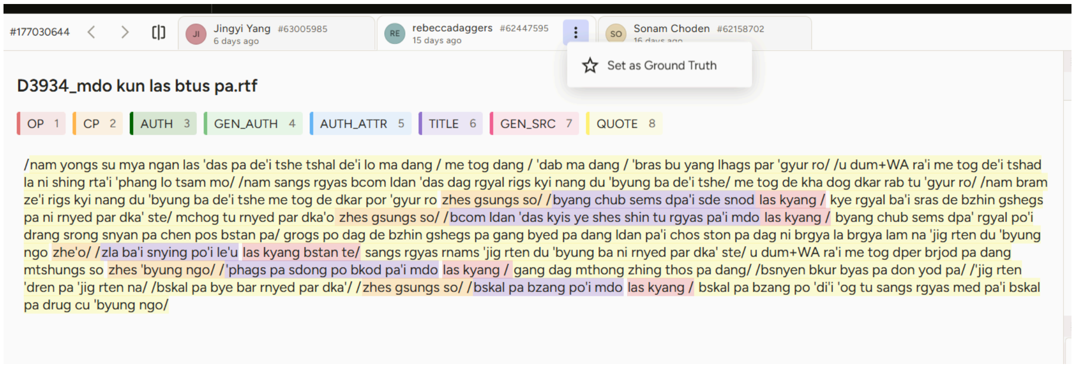


D3934_mdo kun las btus pa.rtf

OP 1 | CP 2 | AUTH 3 | GEN_AUTH 4 | AUTH_ATTR 5 | TITLE 6 | GEN_SRC 7 | QUOTE 8

/nam yongs su mya ngan las 'das pa de'i tshe tshal de'i lo ma dang / me tog dang / 'dab ma dang / 'bras bu yang lhags par 'gyur ro/ /u dum+WA ra'i me tog de'i tshad la ni shing rta'i 'phang lo tsam mo/ /nam sangs rgyas bcom ldan 'das dag rgyal rigs kyi nang du 'byung ba de'i tshe/ me tog de kha dog dkar rab tu 'gyur ro/ /nam bram ze'i rigs kyi nang du 'byung ba de'i tshe me tog de dkar por 'gyur ro zhes gsungs so/ /byang chub sems dpa'i sde snod las kyang / kye rgyal ba'i sras de bzhin gshegs pa ni rnyed par dka' ste/ mchog tu rnyed par dka'o zhes gsungs so/ /bcom ldan 'das kyis ye shes shin tu rgyas pa'i mdo las kyang / byang chub sems dpa' rgyal po'i drang srong snyan pa chen pos bstan pa/ grogs po dag de bzhin gshegs pa gang byed pa dang ldan pa'i chos ston pa dag ni brgya la brgya lam na 'jig rten du 'byung ngo zhe'o/ /zla ba'i snying po'i le'u las kyang bstan te/ sangs rgyas rnams 'jig rten du 'byung ba ni rnyed par dka' ste/ u dum+WA ra'i me tog dper brjod pa dang mtshungs so zhes 'byung ngo/ /'phags pa sdong po bkod pa'i mdo las kyang / gang dag mthong zhing thos pa dang/ /bsnyen bkur byas pa don yod pa/ /'jig rten 'dren pa 'jig rten na/ /bskal pa bye bar rnyed par dka'/ /zhes gsungs so/ /bskal pa bzang po'i mdo las kyang / bskal pa bzang po 'di'i 'og tu sangs rgyas med pa'i bskal pa drug cu 'byung ngo/

1 CP zhes gsungs so/ /
2 QUOTE /nam yongs su mya ngan las
3 TITLE byang chub sems dpa'i sde sn
4 OP las kyang /
5 QUOTE kye rgyal ba'i sras de bzhin g



D3934_mdo kun las btus pa.rtf

/nam yongs su mya ngan las 'das pa de'i tshe tshal de'i lo ma dang / me tog dang / 'dab ma dang / 'bras bu yang lhags par 'gyur ro/ /u dum+WA ra'i me tog de'i tshad la ni shing rta'i 'phang lo tsam mo/ /nam sangs rgyas bcom ldan 'das dag rgyal rigs kyi nang du 'byung ba de'i tshe/ me tog de kha dog dkar rab tu 'gyur ro/ /nam bram ze'i rigs kyi nang du 'byung ba de'i tshe me tog de dkar por 'gyur ro zhes gsungs so/ /byang chub sems dpa'i sde snod las kyang / kye rgyal ba'i sras de bzhin gshegs pa ni rnyed par dka' ste/ mchog tu rnyed par dka'o zhes gsungs so/ /bcom ldan 'das kyis ye shes shin tu rgyas pa'i mdo las kyang / byang chub sems dpa' rgyal po'i drang srong snyan pa chen pos bstan pa/ grogs po dag de bzhin gshegs pa gang byed pa dang ldan pa'i chos ston pa dag ni brgya la brgya lam na 'jig rten du 'byung ngo zhe'o/ /zla ba'i snying po'i le'u las kyang bstan te/ sangs rgyas rnams 'jig rten du 'byung ba ni rnyed par dka' ste/ u dum+WA ra'i me tog dper brjod pa dang mtshungs so zhes 'byung ngo/ /'phags pa sdong po bkod pa'i mdo las kyang / gang dag mthong zhing thos pa dang/ /bsnyen bkur byas pa don yod pa/ /'jig rten 'dren pa 'jig rten na/ /bskal pa bye bar rnyed par dka'/ /zhes gsungs so/ /bskal pa bzang po'i mdo las kyang / bskal pa bzang po 'di'i 'og tu sangs rgyas med pa'i bskal pa drug cu 'byung ngo/