DISAMBIGUATING TIBETAN VERB STEMS WITH MATRIX VERBS IN THE INDIRECT INFINITIVE CONSTRUCTION

EDWARD GARRETT, NATHAN W. HILL, ABEL ZADOKS
SOAS, University of London

Introduction

A great deal of digitized Tibetan texts are now available online, including the entire Derge Kanjur[1] and over 60 Dunhuang documents.[2] Nonetheless, there is a dearth of tools to access and process these data efficiently. To better exploit the profusion of available material, a research project at the School of Oriental and African Studies (SOAS), University of London, is currently working to create a part-of-speech tagged corpus together with an automatic word segmenter and part-of-speech tagger.[3] Garrett et al. (forthcoming a) describes the part-of-speech tag-set used in this project. Garrett et al. (forthcoming b) describes Version 1.0 of the rule-based tagger, designed for Classical Tibetan materials. In the long term, the rule-based tagger will be combined with a statistical tagger to achieve improved results.

Our project began by hand-tagging an initial 17,522 words of the *Mdzaṅs blun*. We developed the initial part-of-speech tag set during this phase. In the next phase, covering the next 26,937 words of the *Mdzaṅs blun* and the first 32,083 words of the *Mi la ras paḥi rnam thar*, we developed the rule-based tagger through an ad hoc process of trial and error. As of 8 June, 2014, the hand-annotated corpus contains 47,927 words of the *Mdzaṅs blun*, 32,083 words of the *Mi la ras paḥi rnam thar*, and 10,069 words of the *Bu ston chos ḥbyuṅ*. The rule-based tagger operates across this corpus with an accuracy of 0.998 and an ambiguity of 1.360.

In addition to the training corpus of hang-tagged data, there is a larger test corpus includes all of these three texts and a number of other

1 Available at http://www.thlib.org/encyclopedias/literary/canons/kt/catalog.php#cat=d on 10 June 2014.

2 Available at otdo.aa.tufs.ac.jp on 10 June 2014.

3 The project 'Tibetan in Digital Communication' is funded by the U.K's Arts and Humanities Research Council.

texts digitized by the University of Otani,[4] including the *Sa-skya legs-bśad, Ḥgro-ba bzaṅ-moḥi rnam-thar, Sa-paṇ-gyi rnam-thar gsuṅ-sgros-ma, Gśen ñi-maḥi rnam-thar, Rwa-loḥi rnam-thar*, and *Mar-paḥi rnam-thar*. This larger corpus permits the testing of hypotheses against a larger data set than the comparatively small hand-tagged training corpus.

The rule-based part-of-speech tagger uses the grammatical rules known to any first year Tibetan student to preclude unlikely part-of-speech interpretations. For example, the tagger knows that the syllable *-so* when it occurs after a verb stem that ends in -s and before the *śad* punctuation mark is not the noun 'tooth'. Unfortunately, for certain verb forms it is not possible in all cases to specify an unambiguous tense analysis.[5] The circumstances giving rise to tense ambiguity are best illustrated with an example. The verb *gśegs* 'go' is invariant across all four tenses. Often syntactic cues disambiguate the correct tense (e.g. *gśegs śig* must be the imperative), but in other contexts disambiguation is not univocal. In the phrase *mi gśegs* the verb *gśegs* is either a present (cf. *mi byed*) or a future (cf. *mi bya*), but cannot be understood as a past. Thus, the tag [v.fut.v.pres] species that in this and comparable contexts it is impractical to decide between [v.fut] and [v.pres]. Finally, there are contexts such as *gśegs śiṅ* and *gśegs so*, in which it is only possible to say that *gśegs* is not the imperative (cf. *byed ciṅ*, *bya źiṅ*, *byas śiṅ*; and *byed do*, *byaḥo*, *byas so*). Rather than tagging such contexts with the lengthy [v.fut.v.past.v.pres] we instead employ the tag [v.invar].[6] Following these protocols the rule-based tagger disambiguates verb stems in those places where they can be disambiguated with certainty, but leaves them ambiguous in cases where the interpretation is not clear-cut.

One of the main purposes of the creation of the corpus and tagger is to yield new insights into Tibetan grammar. Consequently, it would be foolish to rest content with those facets of verb stem distribution well known to the first year Tibetan student. Instead, the tools created so far should be harassed to discover new patters, patterns which in turn can be fed back into refining the corpus and tagger. The current version of

4 Available at http://web.otani.ac.jp/cri/twrp/project/otet/ on 10 June 2014.

5 In this paper the term 'tense' is used to refer to the distinct four principal parts of verbs used in the indigenous grammatical tradition. This terminology is not intended to imply that the morphosyntactic categories recognized by the indigenous tradition correspond semantically to 'tense' (as opposed to 'aspect' or 'mood') as it is used in linguistic typology.

6 One must bear in mind that use of the tag [v.invar] is not a positive claim that a verb is (morphologically or otherwise) invariant, but rather is the negative claim that the stem of this verb in this context cannot be more precisely stated.

the tagger already includes one such new pattern, namely the prohibition of the future stem before the converb *-nas*. Although we introduced this rule on the basis of anecdotal evidence, it appears to be quite robust. For example, all instances of *bya nas* in the Derge Kanjur appear to involve either *bya* 'bird' or *nas* 'barley'. The goal of the current investigation is to see whether verb stems may be disambiguated before the terminative converb in the indirect infinitive construction (e.g. *sloṅ-du ḥoṅs* 'came to ask').

To explore this question we gathered evidence by searching the test corpus for the following types of part-of-speech patterns (cf. appendix).

[v.pres] [cv.term]
[v.fut] [cv.term]
[v.fut.v.pres] [cv.term]
…

We subsequently collated the results of these searches according to the rightward context to list the subcategorization patterns attested with each matrix verb, both in its stem form (e.g. *byed*) and its nominalized form (*byed-pa*).

The subcategorization patterns of positive matrix verbs

This section discusses the various patterns of subcategorization found with unnegated verbs. Verbs appear to fall into three classes, those that take future and present stems with equal preference, those that select for the present, and those that select for the future. Although the corpus does provide examples of subordinated past stems, this phenomenon appears to be unsystematic.

Eight matrix verbs occur with subordinate present and future stems and with stems that are ambiguous between present and future.

bcug(*-pa*) (24 v.pres, 20 v.fut, 6 v.fut.v.pres), *gźug*(*-pa*) (5 v.pres, 1 v.fut, 4 v.fut.v.pres), *ḥjug* (3 v.pres, 1 v.fut), *chug* (1 v.pres, 1 v.fut.v.pres)
soṅ(*-ba*) (14 v.pres, 4 v.fut, 12 v.fut.v.pres), *ḥgro* (11 v.pres, 2 v.fut, 1 v.fut.v.pres), *phyin*(*-pa*) (9 v.pres, 3 v.fut, 2 v.fut.v.pres)
btaṅ(*-ba*) (20 v.pres, 8 v.fut, 12 v.fut.v.pres), *gtoṅ* (1 v.pres, 1 v.fut.v.pres)
byon(*-pa*) (5 v.pres, 5 v.fut, 10 v.fut.v.pres), *ḥbyon-pa* (1 v.pres)
gyur(*-pa*/*-ba*) (3 v.pres, 3 v.fut, 9 v.fut.v.pres), *ḥgyur* (3

v.fut.v.pres)
byuṅ (2 v.pres, 1 v.fut, 8 v.fut.v.pres)
doṅ-ba (1 v.pres, 1 v.fut, 1 v.fut.v.pres), *ḥdoṅ* (1 v.fut.v.pres)
smra-ba (1 v.pres, 1 v.fut), *smras* (1 v.fut.v.pres)

Six matrix verbs appear to select only for the present stem. In the corpus there are examples of these matrix verbs selecting presents and selecting verb stems that are ambiguous between present and future, but there are no examples of these six verb selecting unambiguous future stems.

ḥoṅs(*-pa*) (12 v.pres, 7 v.fut.v.pres), *ḥoṅ*(*-ba)* (11 v.pres, 5 v.fut.v.pres), *yoṅ-ba* (4 v.fut.v.pres), *yoṅs-pa* (1 v.pres, 1 v.fut.v.pres)
mdzad(*-pa*) (1 v.pres, 13 v.fut.v.pres)
byas(*-pa*) (2 v.pres, 2 v.fut.v.pres)
btags-pa (1 v.pres, 2 v.fut.v.pres)
bzuṅ-pa (1 v.pres, 1 v.fut.v.pres), *ḥdzin-pa* (1 v.fut.v.pres)
mchi (1 v.pres, 1 v.fut.v.pres)

One must however bear in mind that it is always possible the absence of unambiguous future stems in the subordinate clause might be an accident gap in the corpus, rather than a structural fact about the Tibetan language. It is only the verb *ḥoṅ/yoṅs* which is well attested enough that it seems quite likely that the occurrence of future stems in the subordinate clause can be precluded.

There are also six matrix verbs that appear to select only the future stem. In the corpus there are examples of these matrix verbs selecting futures and selecting verb stems that are ambiguous between future and present, but there are no examples of these six verb selecting unambiguous present stems.

gsol (19 v.fut, 7 v.fut.v.pres)
med(*-pa*) (7 v.fut, 8 v.fut.v.pres)
grags(*-pa*) (2 v.fut, 1 v.fut.v.pres)
yod-pa (2 v.fut, 2 v.fut.v.pres)
ruṅ-ba (1 v.fut, 1 v.fut.v.pres)
gśegs-pa (1 v.fut, 1 v.fut.v.pres)

The fact that these six verbs show no stem changes lends credence to their interpretation as a structural class. The presence of both *med* and *yod* in this category cannot be a coincidence. In contrast, in light of the fact that motion verbs general appear among those verbs that take the present and future with equal preference, it seems likely that a lack of *gśegs* selecting a present stem is an accidental gap.

Turning to the past stems in subordinate clauses, there are some surprises. Although past stems are the most common stem in the corpus overall, only four verbs occur more than once as matrix verbs selecting the past stem: 1. *gsol*(*-ba*) (9), 2. *soṅ* (2), *phyin* (1), *ḥgro-ba* (1), 3. *ḥoṅ* (2), *ḥoṅs-pa* (1), 4. *yod*-(*pa*) (3). All four have been discussed previously. The first (*gsol*) and fourth (*yod*) generally selects for the future. The second (*ḥgro/phyin/soṅ*) selects for either present or future. The third (*ḥoṅ*) selects the present. If one looks at the actual occurrences of these four verbs when they appear to select the past stem, the past stems in question turn out to predominantly to be misspellings of the future stem.

bzuṅ (*gzuṅ*) *du gsol* (5)
bcad (*gcad*) *du gsol* (1)
bcad (*gcad*) *du ḥoṅs-pa* (1)
bsad (*gsad*) *du ḥoṅ* (1)
byuṅ (*ḥbyuṅ*) *du soṅ* (1)
byuṅ (*ḥbyuṅ*) *du ḥgro-ba* (1)
byuṅ (*ḥbyuṅ*) *du yod-pa* (2)

This misspelling of futures as pasts is not unexpected, since the spoken languages lose the future and often the etymological future is homophonous in a given dialect with an etymological past because of sound change. In some cases the sandhi makes clear that a future was intended.

gyur du (past requires -*tu*) *ḥoṅ* (1)
bsgrubs tu (past requires -*su*) *gsol-ba* (1)
bsgrubs tu (past requires -*su*) *yod* (1)

All of these cases merit more detailed philological discussion; here it suffices to conclude that there is no solid evidence for a class of matrix verbs that select the past stem in the subordinate clause.

For the purposes of verb stem disambiguation, one may conclude that if a verb has any interpretation other than [v.past] and it occurs as the subordinate verb of this infinitive construction, then the analysis of this verb as [v.past] may be removed. One may more tentatively conclude that a verb that has interpretations other than [v.fut] (and [v.past]) should have [v.fut] removed, if the matrix verb is one of the six verbs that selects the present, although perhaps this stipulation is only safely applied to *ḥoṅ* and its forms. Because there are few futures in the corpus in general the lack of a future before the other verbs seven present selecting verbs may be a coincidence.

Reciprocally, if the matrix verb is one of the six future selecting verbs, then the analysis of the subordinate verb as [v.pres] can be precluded. In this case, the overall prevalence of presents over futures in the corpus means that the lack of presents before these verbs can be safely understood as a systematic gap. Furthermore, the fact that all future selecting verbs are invariant in their stem morphology, and that *yod* and *med* both occur in this category, bolsters the security of the analysis. Nonetheless, because *gśegs* is the only motion verb in this group, and most motion verbs can select either present or future stem, it would be a mistake to forbid the analysis of a verb subordinate to *gśegs* as a present.

The subcategorization patterns of negative matrix verbs

Because negated verbs are less frequent than positive verbs, the evidence of negated matrix verbs in the indirect infinitive construction is sparse. However, the investigation of subcategorization patterns with positive verbs provides a framework to analyze these data, i.e. if the evidence of negated verbs confirms the patterns seen with the unnegated verbs then the pattern in question more secure.

Four of the eight verbs that select for both present and future are attested with negation.

ḥjug-pa (1 v.pres), *bcug*(*-pa*) (2 v.pres, 1 v.fut.v.pres)
byuṅ (1 v.pres, 1 v.fut)
phyin (1 v.fut.v.pres)
gtoṅ-ba (1 v.fut.v.pres)

These data support the perspective that these verbs select both present and future, although they suffice to conclude this only in the cases of *byuṅ*.

Turning to the verbs that in the positive appear to select for the present, the evidence from negated matrix verbs is very thin.

byas (1 v.pres)
yoṅ (1 v.fut), *ḥoṅs* (1 v.fut.v.pres), *ḥoṅ* (1 v.fut.v.pres)

The occurrence of *yoṅ* with an unambiguous subordinate future weighs against the theory that this verb selects for the present. The deviant examples is *bsgrub tu ma yoṅ-ba* 'did not come to accomplish'. One may perhaps suspect a spelling mistake of *bsgrub* for *sgrub*.

The verb *ruṅ* is the only verb that selects for future for which negated examples occur in the corpus. In fact, this verb occurs far more frequently negated than in the positive form. Most of the negated forms

support the analysis of this verb as selecting a future stem.

ruṅ (11 v.fut, 4 v.fut.v.pres, 2 v.fut.v.past)

However, there are two problem cases. In one of these cases *ruṅ* selects the past stem *byuṅ*, which is possibly a simple misspelling for *ḥbyuṅ.* The other case is a genuine problem. The passage is *gcod du mi ruṅ* with an unambiguous present. So, paradoxically although (*mi*) *ruṅ* selecting the future is among the most robust patterns in the corpus, *gcod du mi ruṅ* is also one of the most clear-cut counter-examples to any pattern identifiable in the corpus.

A number of matrix verbs that predominantly appear in the negative were not noted in the discussion of positive matrix verbs. The verb *chud* appears to only select for the present.

chud (1 v.pres, 1 non-negated v.pres, 7 v.fut.v.pres)

Three verbs appear to select the present and future with equal ease.

btub(*-pa*) (2 v.pres, 1 v.fut, 10 v.fut.v.pres), *gtub-pa* (1 v.pres)
ḥdod(*-pa*) (1 v.pres, 3 v.fut.v.pres, 1 v.fut.v.past)
snaṅ (1 v.pres, 1 non-negated v.fut)

Concerning the verb *tshud*, one can only say that it does not select the past, which is not surprising since no verbs appear to select the past.

tshud (3 v.fut.v.pres)

The verb *gnaṅ* has a strange pattern of subcategorization, appearing to select the past in two cases (once selecting *phyuṅ* and once *byuṅ*).

gnaṅ (1 v.fut, 2 v.past)

The verbs *snaṅ*, *tshud*, and *gnaṅ* are not well enough attested in the corpus to yield any reliable conclusion.

Conclusions

With a certain fuzziness around the edges, the data surveyed here suggest that past tense verbs do not occur as the subordinate verbs of indirect infinitives and that the matrix verbs *gsol, med, grags, yod, ruṅ* select the future tense in the subordinate clause. It is possible that one group of verbs selects the present tense whereas others are equally happy to select the present and the future, but the overall rarity of future stems in the corpus makes the line between these two categories difficult to draw.

REFERENCES

Garrett, Edward, Nathan W. Hill, Adam Kilgarif, Ravikiran Vadlapudi, Abel Zadoks (forthcoming a). "The contribution of corpus linguistics to lexicography and the future of Tibetan dictionaries." *Revue d'Etudes Tibétaines*.

Garrett, Edward, Nathan W. Hill, Abel Zadoks (forthcoming b). "A Rule-based Part-of-speech Tagger for Classical Tibetan." *Himalayan Linguistics*.

APPENDIX: MATRIX VERBS IN INDIRECT INFINITIVE CONSTRUCTION

[v.pres] [cv.term]

bcug(*-pa*) (24), *gźug*(*-pa*) (5), *ḥjug* (3), *chug* (1)
soṅ(*-ba*) (14), *ḥgro* (11), *phyin*(*-pa*) (9)
ḥoṅs(*-pa*) (12), *ḥoṅ*(*-ba*) (11), *yoṅs-pa* (1)
btaṅ(*-ba*) (20), *gtoṅ* (1)
byon(*-pa*) (5), *ḥbyon-pa* (1)
gyur(*-pa*) (3)
ḥchad(*-pa*) (2), *bśad* (1)
tshor-ba (2)
bsgyur-pa (1), *bsgyur-ba* (1)
byuṅ (2)
byas(*-pa*) (2)
byiṅ(*-ba*) (2)
chud (1)
mchi (1)
gñag-ba (1)
btags-pa (1)
doṅ-ba (1)
draṅs-pa (1)
smra-ba (1)
btsugs (1)
brtsams-pa (1)
brtsis-pa (1)
mdzad (1)
brdzaṅs-pa (1)
bzuṅ-pa (1)
len (1)

[v.fut] [cv.term]

bcug-pa (20), *gźug-pa* (1), *ḥjug-pa* (1)
gsol (19)
soṅ(*-ba*) (4), *phyin-pa* (3), *ḥgro* (2)
btaṅ(*-ba*) (8)
med(*-pa*) (7)
byon(*-pa*) (5)
gyur(*-ba*) (3)
bsdus-pa (2)
grags(*-pa*) (2)
yod-pa (2)
doṅ-ba (1)
nus (1)
snaṅ (1)
byuṅ (1)
smra (1)
źugs (1)
ruṅ-ba (1)
gśegs-pa (1)

[v.fut.v.pres] [cv.term]

soṅ (12), *phyin* (2), *ḥgro* (1)
btaṅ(*-pa*) (12), *gtoṅ* (1)
gyur(*-pa*) (9), *ḥgyur* (3)
ḥoṅs(*-pa*) (7), *ḥoṅ*(*-ba*) (5), *yoṅ-*

ba (4), *yoṅs-pa* (1)
bcug(-*pa*) (6), *gźug*(-*pa*) (4), *chug* (1)
mdzad(-*pa*) (13)
byon(-*pa*) (10)
gsol (8)
byuṅ (8)
med(-*pa*) (8)
bkod(-*pa*) (4), *ḥgod-pa* (1)
cha-ba (3), *chas-pa* (1)
phebs(-*pa*) (3)
bzuṅ (1), *ḥdzin-pa* (1)
doṅ-ba (1), *ḥdoṅ* (1)
khrid-pa (2), *ḥkhrid* (1)
mkhyen (2)
ñal (2)
btags (2)
byas (2)
yod-pa (2)
bskyaṅs (1)
khol-ba (1)
grags (1)
rgyas (1)
bgyi-ba (1)
rgyugs-pa (1)
ṅes (1)
mchi (1)
ḥchos-pa (1)
ñe (1)
rtogs-pa (1)
mthoṅ (1)
btub (1)
ltem (1)
dod (1)
phul (1)
phed (1)
ḥbab (1)
smras (1)
btsal (1)
stsol-ba (1)
ḥdzud (1)
u (1)
bzo (1)
ruṅ-ba (1)
śes-pa (1)
gśegs (1)
lhags-pa (1)
lhuṅ-ba (1)

[v.past] [cv.term]

gsol(-*ba*) (9)
soṅ (2)
ḥoṅ (2), *ḥoṅs*-pa (1)
yod-(pa) (3)
mid-pa (1)
ñe-ba (1)
lhuṅ-ba (1)
byin-pa (1)
bsdus-pa (1)
byas-pa (1)
gsuṅs-pa (1)
byon-pa (1)
skye-ba (1)
snyoms-pa (1)
mṅaḥ (1)
ḥgril (1)
bcad (1)
sbyar (1)
bcug (1)
ḥkhruṅs (1)
spyod (1)

[v.fut.v.past] [cv.term]

gsol (11)
med(-*pa*) (2)
yoṅ (1), *yoṅs-pa* (1)
śog (2)

btuṅ (1)
byuṅ (1)
bźugs (1)

[v.pres] [cv.term] [neg]

ḥjug-pa (1), *bcug*(*-pa*) (2)
btub (2), *gtub-pa* (1)
ruṅ (1)
dbaṅ (1)
ḥdod (1)
byas (1)
byuṅ (1)
snaṅ (1)
chud (1)
ster-ba (1)

[v.fut] [cv.term] [neg]

ruṅ (11)
gnaṅ (1)
btub (1)
bzad-pa (1)
zad-pa (1)
yoṅ (1)
byuṅ (1)

[v.fut.v.pres] [cv.term] [neg]

btub(**-***pa*) (10)
chud (7)
ruṅ (4)
tshud (3)
ḥdod(**-***pa*) (3)
ḥoṅs (1), *ḥoṅ* (1)
bcug-pa (1)
phyin (1)
e (1)
stsal (1)
*dad***-***pa* (1)
gtoṅ-ba (1)

[v.past] [cv.term] [neg]

gnaṅ (2)
ruṅ (1)
ltuṅ-ba (1)

[v.fut.v.past] [cv.term] [neg]

ruṅ (2)
re (1)
ḥdod-pa (1)