A rule based tagger for Classical Tibetan:

negation and verb stems classification

Nathan W. Hill and Edward Garrett

(SOAS, University of London)

Tibetan in Digital Communication

Goals

1. 1,000,000 word part-of-speech tagged corpus of Tibetan texts
2. An automatic word breaker
3. An automatic part-of-speech tagger

Tibetan in Digital Communication

Goals

1. 1,000,000 word part-of-speech tagged corpus of Tibetan texts
2. An automatic word breaker
3. An automatic part-of-speech tagger

Here we focus on the development of rule-based part-of-speech tagging.

Tibetan in Digital Communication

Goals

1. 1,000,000 word part-of-speech tagged corpus of Tibetan texts
2. An automatic word breaker
3. An automatic part-of-speech tagger

Here we focus on the development of rule-based part-of-speech tagging.

1. An overview of the project workflow

Tibetan in Digital Communication

Goals

1. 1,000,000 word part-of-speech tagged corpus of Tibetan texts
2. An automatic word breaker
3. An automatic part-of-speech tagger

Here we focus on the development of rule-based part-of-speech tagging.

1. An overview of the project workflow
2. A simple example of a rule used by the tagger (*so* ‘tooth’ or verb suffix)

Tibetan in Digital Communication

Goals

1. 1,000,000 word part-of-speech tagged corpus of Tibetan texts
2. An automatic word breaker
3. An automatic part-of-speech tagger

Here we focus on the development of rule-based part-of-speech tagging.

1. An overview of the project workflow
2. A simple example of a rule used by the tagger (*so* 'tooth' or verb suffix)
3. Distinguishing *mi* and *ma* as nouns and negation markers

Tibetan in Digital Communication

Goals

1. 1,000,000 word part-of-speech tagged corpus of Tibetan texts
2. An automatic word breaker
3. An automatic part-of-speech tagger

Here we focus on the development of rule-based part-of-speech tagging.

1. An overview of the project workflow
2. A simple example of a rule used by the tagger (*so* 'tooth' or verb suffix)
3. Distinguishing *mi* and *ma* as nouns and negation markers
4. Disambiguating verb stems (main focus of the presentation)

Workflow:
(1) Look-up of possible analyses

Word	Part-of-speech tag
ཆུལ་པོ་	n.count
ཤེ་	d.dem ~ cv.sem
ལ་	case.all ~ n.count
བཙུན་མོ་	n.count
ཉ་	num.card
བཅུ་	num.card
ཡོད་	v.invar
ཀྱང་	cl.focus
།	punc

Workflow:
(2) Pre-tagging

Word	Part-of-speech tag
ཆུལ་པོ་	n.count
ཁ་	d.dem
ཡ་	case.all ~ n.count
བཙུན་མོ་	n.count
ཉི་	num.card
བཅུ་	num.card
ཡོད་	v.invar
ཀྱང་	cl.focus
།	punc

Workflow:
(3) Hand-tagging

Word	Part-of-speech tag
ཆུལ་ཡོ་	n.count
ཤ་	d.dem
ཡ་	case.all
བཙུན་མོ་	n.count
ཐ་	num.card
བཅུ་	num.card
ཡོད་	v.invar
ཀྱང་	cl.focus
།	punc

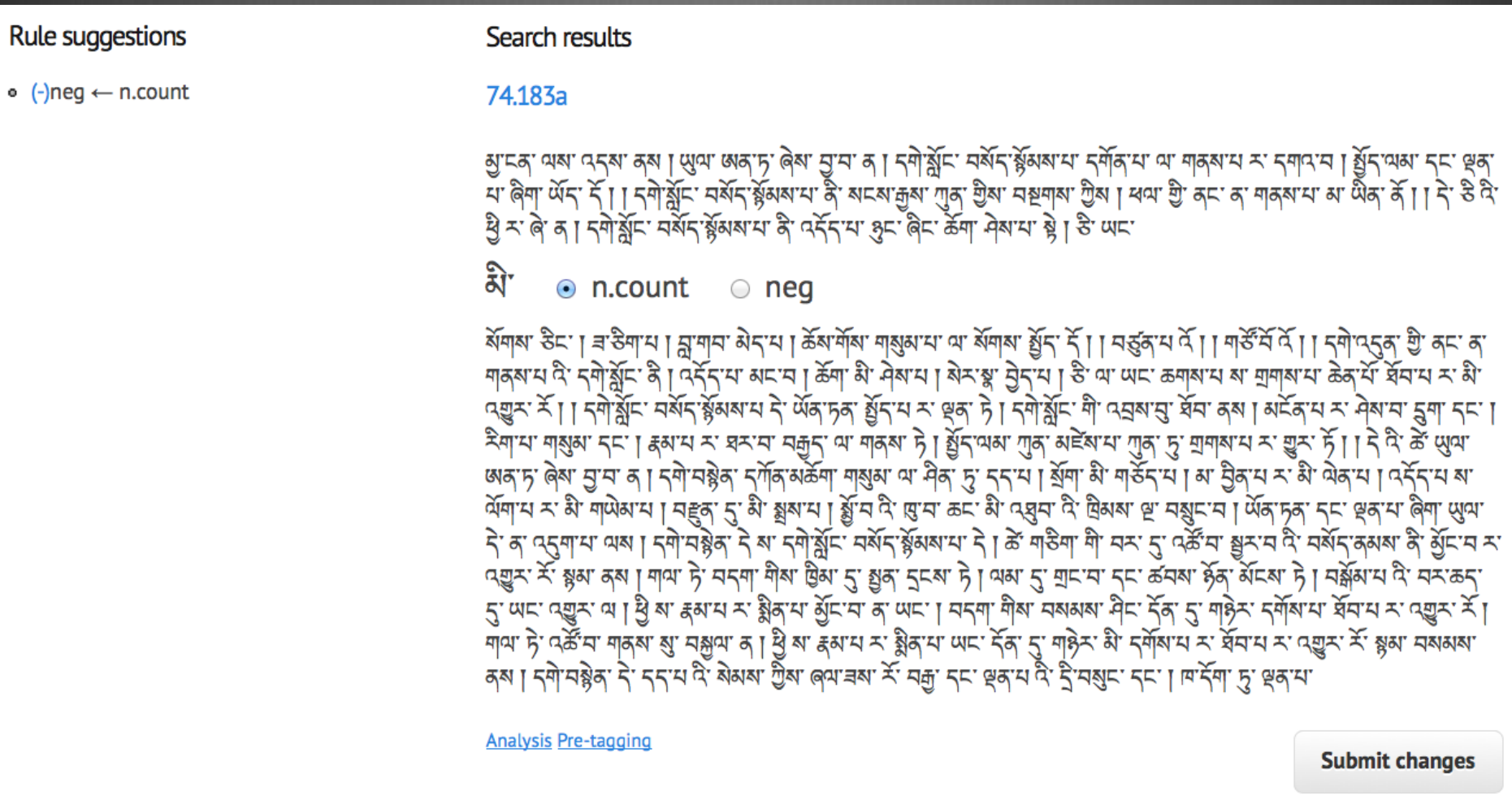
Workflow:
(4) Rule suggestions

Screen shot of rule suggestions

(9 November 2013)

Workflow:
(4) Rule suggestions

Screen shot of the rule suggestion [neg] ← [n.count] (9 November 2013)



A simple example: Identifying sandhi determined suffixes

so 'tooth' versus *so* finite verb suffix

A simple example: Identifying sandhi determined suffixes

so 'tooth' versus *so* finite verb suffix

finite verb suffix always after -s, but

A simple example: Identifying sandhi determined suffixes

so 'tooth' versus *so* finite verb suffix

finite verb suffix always after -s, but

khos so bcag 'he broke teeth'

A simple example: Identifying sandhi determined suffixes

so 'tooth' versus *so* finite verb suffix

finite verb suffix always after -s, but

khos so bcag 'he broke teeth'

will yield false positive

A simple example: Identifying sandhi determined suffixes

so 'tooth' versus *so* finite verb suffix

finite verb suffix always after -s, but

khos so bcag 'he broke teeth'

will yield false positive

So, we also specify that after *so* is *śad*, or the syllables *źes*, *sñam*, or *zer*.

A simple example: Identifying sandhi determined suffixes

so 'tooth' versus *so* finite verb suffix

finite verb suffix always after -s, but

khos so bcag 'he broke teeth'

will yield false positive

So, we also specify that after *so* is *śad*, or the syllables *źes*, *sñam*, or *zer*.

Similar rules for *-lo*, *-ro*, *-de*, *-śiṅ,* etc.

Disambiguating *mi* and *ma* as negations or nouns

(1). Finding the nouns *mi and ma* within noun phrases

mi ḥdi 'this person'

mi gsum 'three people'

ma rgan-mo 'your old mother'

Disambiguating *mi* and *ma* as negations or nouns

(2). Isolating *mi* [n.count] and *ma* [n.count] after the genitive

rmoṅ-pa ḥi mi ḥgro ḥo
'an ignorant person goes'.

Disambiguating *mi* and *ma* as negations or nouns

(2). Isolating *mi* [n.count] and *ma* [n.count] after the genitive

rmoṅ-pa ḥi mi ḥgro ḥo
'an ignorant person goes'.

bskal-pa graṅs med-pa ḥi mi dge-ba ḥi las
'non virtuous deeds of countless eons'.

Disambiguating *mi* and *ma* as negations or nouns

(2). Isolating *mi* [n.count] and *ma* [n.count] after the genitive

rmoṅ-pa ḥi mi ḥgro ḥo
'an ignorant person goes'.

bskal-pa graṅs med-pa ḥi mi dge-ba ḥi las
'non virtuous deeds of countless eons'.

rab tu ḥbyuṅ-ba ḥi mi rigs
'it is not proper to take ordination'.

Disambiguating *mi* and *ma* as negations or nouns

(2). Isolating *mi* [n.count] and *ma* [n.count] after the genitive

rmoṅ-pa ḥi mi ḥgro ḥo
'an ignorant person goes'.

bskal-pa graṅs med-pa ḥi mi dge-ba ḥi las
'non virtuous deeds of countless eons'.

rab tu ḥbyuṅ-ba ḥi mi rigs
'it is not proper to take ordination'.

RULE: If *mi / ma* could be [n.count], follows a probable genitive, does not precede *rigs*, and does not precede a [n.v.xxx], and the word before the probable genitive is not an unambiguous [v.xxx] tag, then mark *mi / ma* as a [n.count].

Disambiguating *mi* and *ma* as negations or nouns

(2). Isolating *mi* [n.count] and *ma* [n.count] after the genitive

rmoṅ-pa ḥi mi ḥgro ḥo
'an ignorant person goes'.

bskal-pa graṅs med-pa ḥi mi dge-ba ḥi las
'non virtuous deeds of countless eons'.

rab tu ḥbyuṅ-ba ḥi mi rigs
'it is not proper to take ordination'.

RULE: If *mi / ma* could be [n.count], follows a probable genitive, does not precede *rigs*, and does not precede a [n.v.xxx], and the word before the probable genitive is not an unambiguous [v.xxx] tag, then mark *mi / ma* as a [n.count].

PATTERN: (\S+\|(?:\[(?!v\.)[^\]]*\])+\s+(?:འི་|ཀྱི་|གི་|ཡི་)\|\S+\s+(?:མི་|མ་))\|\S*\[n\.count\]\S*(?!\s+(?:རིགས་\||\S+\[n\.v\.))

REPLACE: $1|[n.count]

Disambiguating *mi* and *ma* as negations or nouns

(3). Isolating *mi* and *ma* as nouns before *med-pa*

mi med-paḥi sa 'a place without people'

bu ma med-pa 'a boy without a mother'

Disambiguating *mi* and *ma* as negations or nouns

(4). Using the associative case to find nouns

lha daṅ mi 'gods and men'

Disambiguating *mi* and *ma* as negations or nouns

(5). Identifying *mi* [neg] before present and future tense verb forms

mi byed do 'doesn't do'

mi gśegs so 'doesn't go'

Disambiguating *mi* and *ma* as negations or nouns

(5). Identifying *mi* [neg] before present and future tense verb forms

mi byed do 'doesn't do'

mi gśegs so 'doesn't go'

mi gaṅ 'not full' and 'which person'

Disambiguating *mi* and *ma* as negations or nouns

(5). Identifying *mi* [neg] before present and future tense verb forms

mi byed do 'doesn't do'

mi gśegs so 'doesn't go'

mi gaṅ 'not full' and 'which person'

mi sogs te 'people etc’

Disambiguating *mi* and *ma* as negations or nouns

(5). Identifying *mi* [neg] before present and future tense verb forms

mi byed do 'doesn't do'

mi gśegs so 'doesn't go'

mi gaṅ 'not full' and 'which person'

mi sogs te 'people etc'

mi bgres-pa 'aged person'

Disambiguating *mi* and *ma* as negations or nouns

(6). Identifying *ma* [neg] in the prohibitive

ma byed 'do not do it!'

ma gśegs śig 'do not go!'

Disambiguating *mi* and *ma* as negations or nouns

(7). Isolating *ma* [neg] before the past tense and *yin*

ma byas 'didn't do it'

ma gśegs so 'didn't go'

ma yin te 'isn't' (rarely 'is a mother')

Distinguishing the four tenses and subsequent cleanup

As an default hypothesis, we follow the dictionaries (cf. Hill 2010) in assuming that all Tibetan verbs in principle distinguish four tenses, the present [v.pres], the past [v.past], the future [v.fut], and the imperative [v.imp].

Disambiguating verb tenses

1. Morphology
2. Syntax
3. *da-drag* sandhi

Morphology

No problem when all four tenses are distinct.

pres. *byed*, past *byas*, fut. *bya*, imp. *byos* 'do'

pres. *gsod*, past *bsad*, fut. *gsad*, imp. *sod* 'kill'

Morphology

No problem when all four tenses are distinct.

pres. *byed*, past *byas*, fut. *bya*, imp. *byos* 'do'

pres. *gsod*, past *bsad*, fut. *gsad*, imp. *sod* 'kill'

A problem when all four stems are not distinct.

pres. *gśegs*, past *gśegs*, fut. *gśegs*, imp. *gśegs* 'go'

pres. *mdzad*, past *mdzad*, fut. *mdzad*, imp. *mdzod* 'do'

pres. *skyel*, past *bskyal*, fut. *bskyal*, imp. *skyol* 'carry'

Syntax:
Using verbal suffixes

(1). Finding the imperative before [cv.imp]

byos śig 'do it!' (imperative)

gśegs śig 'go!' (imperative)

Syntax:
Using verbal suffixes

(1). Finding the imperative before [cv.imp]

byos śig 'do it!' (imperative)

gśegs śig 'go!' (imperative)

But watch out

ma gśegs śig 'do not go!' (present)

gnas śig 'stay!' (imperative) or 'a place' (noun)

Syntax:
Using verbal suffixes

(2). Finding the prohibitive (present negated with *ma*) before [cv.imp]

ma byed cig 'do not do!' (present)

ma gśegs śig 'do not go!' (present)

Syntax:
Using verbal suffixes

(3). Prohibiting the imperative in non-finite contexts

(None of the following are the imperative.)

gśegs na 'if he goes'

gśegs kyaṅ 'although he goes'

gśegs nas 'after going'

gśegs śiṅ 'goes and … '

gśegs te 'going… '

Syntax:
Using verbal suffixes

(3b). Prohibiting the imperative in explicitly non-imperative, but finite contexts

(Neither of the following is an imperative.)

gśegs sam 'is he going?'

gśegs so 'he goes'

Syntax:
Using verbal suffixes

(4). The prohibition of the future before the elative converb *nas*

byed nas 'after doing' (present)

byas nas 'after doing' (past)

*bya nas (unattested)

gśegs nas is not future (present or past)

Syntax:
Using negation

(5). Negation with *ma* implies past

ma byas so 'didn't do' (past)

ma gśegs so 'didn't go' (past)

Syntax:
Using negation

(6). Negation with *mi* precludes past

mi byed 'isn't doing' (present)

mi bya 'won't do' (future)

mi gśegs 'isn't going, won't go' (present or future)

(N.B. *mi gśegs nas* present only)

da-drag sandhi

Future	**Past**
bskyal lo	*bskyal to*
bskyal lam	*bskyal tam*
bskyal du	*bskyal tu*
bskyal yaṅ	*bskyal kyaṅ*
bskyal ba	*bskyal pa*

da-drag sandhi

Not all past forms have a *da-drag*.

'give' (*sbyin, byin, sbyin, byin*), always *byin no* (not *byin to*).

Some *da-drag* forms are presents.

'take' (*ḥdzin, bzuṅ, gzuṅ, zuṅs*), always *ḥdzin to* (not *ḥdzin no*).

da-drag sandhi

Not all past forms have a *da-drag*.

'give' (*sbyin, byin, sbyin, byin*), always *byin no* (not *byin to*).

Some *da-drag* forms are presents.

'take' (*ḥdzin, bzuṅ, gzuṅ, zuṅs*), always *ḥdzin to* (not *ḥdzin no*).

Because *-pa* and *-ba* are similar looking and frequently confused, this rule may seem to risk introducing errors. However, we think it is best to disambiguate verb stems wherever it is possible to do so. Disambiguating these stems permits the behavior of *-pa* versus *-ba* to be more easily explored by future researchers; reason enough to add the rule.

The correct ordering of disambiguation strategies

1.

gaṅ źig śin tu dad-pa ḥi sems kyis chu sñam-pa gaṅ tsam saṅs-rgyas la mchod dam ˩ dge ḥdun la phul [v.past] ~ [v.imp] *lam ˩ pha-ma la phul* [v.past] ~ [v.imp] *lam ˩ dbul-phoṅs la byin nas ˩ gcan-gzan la byin na ˩ gsod-nams ḥdi ni bskal-ba du-ma r yaṅ mi zad de ˩*

"If one with a mind of great faith offers handfuls of water to the buddha, or makes offerings to the saṃgha, or makes offerings to one's parents, or gives to the poor, or gives to wild animals, this merit shall not run out for many eons".

(pres. *ḥbul*, past *phul*, fut. *dbul*, imp. *phul*) 'give'

The correct ordering of disambiguation strategies

1.

gaṅ źig śin tu dad-pa ḥi sems kyis chu sñam-pa gaṅ tsam saṅs-rgyas la mchod dam ˩ dge ḥdun la phul [v.past] ~ [v.imp] *lam ˩ pha-ma la phul* [v.past] ~ [v.imp] *lam ˩ dbul-phoṅs la byin nas ˩ gcan-gzan la byin na ˩ gsod-nams ḥdi ni bskal-ba du-ma r yaṅ mi zad de ˩*

"If one with a mind of great faith offers handfuls of water to the buddha, or makes offerings to the saṃgha, or makes offerings to one's parents, or gives to the poor, or gives to wild animals, this merit shall not run out for many eons".

(pres. *ḥbul*, past *phul*, fut. *dbul*, imp. *phul*) 'give'

If *da-drag* rules are ordered before the exlusion of [v.imp] before [cv.ques], then the machine concludes that *phul* is [v.imp], since the suffix should have been *–tam* for the past.

The correct ordering of disambiguation strategies

2.

ku-śu ḥdi ni ḥbras-bu las skyes-pa ma lags te ⌋ chab-mig cig gi naṅ nas rñed-pa s slan-cad ni bdag gis mi rñed de ⌋ mi ḥbyor [v.past] ~ [v.pres] *to*

"This apple was not born from fruit, but I found it from inside a spring, so I cannot find it hereafter. It will not be encountered."

The correct ordering of disambiguation strategies

2.

ku-śu ḥdi ni ḥbras-bu las skyes-pa ma lags te ⌋ chab-mig cig gi naṅ nas rñed-pa s slan-cad ni bdag gis mi rñed de ⌋ mi ḥbyor [v.past] ~ [v.pres] *to*

"This apple was not born from fruit, but I found it from inside a spring, so I cannot find it hereafter. It will not be encountered."

If *da-drag* rules are ordered before the rule that prevents the past stem after *mi,* then the suffix -*to* would have triggered [v.past].

The correct ordering of disambiguation strategies

2.

ku-śu ḥdi ni ḥbras-bu las skyes-pa ma lags te ⌋ chab-mig cig gi naṅ nas rñed-pa s slan-cad ni bdag gis mi rñed de ⌋ mi ḥbyor [v.past] ~ [v.pres] *to*

"This apple was not born from fruit, but I found it from inside a spring, so I cannot find it hereafter. It will not be encountered."

If *da-drag* rules are ordered before the rule that prevents the past stem after *mi,* then the suffix -*to* would have triggered [v.past].

We order negation rules before *da-drag* rules, so *ḥbyor* is tagged [v.pres].

The correct ordering of disambiguation strategies

3.

gal te sñiṅ nas ma btsal [v.past] ~ [v.fut] *lam le-lo źig byas te* ‖ …

'if one does not seek wholeheartedly, or is lazy… '.

Preliminary observations on the distribution of *da-drag*

1. *zer* 'say' never gets a *da-drag*
2. *bden* 'be true',

    as a finite verb always *bden no* (without *da-drag*)

    as a verbal noun always *bden-pa* (with *da-drag*)

We only go so far

bdag rab tu dbyuṅ du gsol
'I request that you give me ordination'.

bdag la saṅs-rgyas kyi chos bśad du gsol
'I request that you explain to me the Buddha's dharma'.

We only go so far

bdag rab tu dbyuṅ du gsol
'I request that you give me ordination'.

bdag la saṅs-rgyas kyi chos bśad du gsol
'I request that you explain to me the Buddha's dharma'.

One could argue that since *dbyuṅ* is a future, then *bśad* is also a future.

We only go so far

bdag rab tu dbyuṅ du gsol
'I request that you give me ordination'.

bdag la saṅs-rgyas kyi chos bśad du gsol
'I request that you explain to me the Buddha's dharma'.

One could argue that since *dbyuṅ* is a future, then *bśad* is also a future.

But this is debatable, and hard for a computer to notice.

Consolidating ambiguous verb forms into ambiguous tags

bskyed [v.fut] ~ [v.past] *do* [cv.fin]

Consolidating ambiguous verb forms into ambiguous tags

bskyed [v.fut] ~ [v.past] *do* [cv.fin]

> *bskyed* [v.fut.v.past] *do* [cv.fin]

Consolidating ambiguous verb forms into ambiguous tags

bskyed [v.fut] ~ [v.past] *do* [cv.fin]

> *bskyed* [v.fut.v.past] *do* [cv.fin]

gśegs [v.fut] ~ [v.past] ~ [v.pres] *so* [cv.fin]

Consolidating ambiguous verb forms into ambiguous tags

bskyed [v.fut] ~ [v.past] *do* [cv.fin]

> *bskyed* [v.fut.v.past] *do* [cv.fin]

gśegs [v.fut] ~ [v.past] ~ [v.pres] *so* [cv.fin]

> *gśegs* [v.invar] *so* [cv.fin]

Restoring ambiguity when a single form might belong to two distinct verbs

The preceding rules would make the change

źu [v.fut] ~ [v.past] ~ [v.pres]

Restoring ambiguity when a single form might belong to two distinct verbs

The preceding rules would make the change

źu [v.fut] ~ [v.past] ~ [v.pres]

> *źu* [v.invar]

Restoring ambiguity when a single form might belong to two distinct verbs

The preceding rules would make the change

źu [v.fut] ~ [v.past] ~ [v.pres]

> *źu* [v.invar]

but *źu* [v.fut] ~ [v.pres] belong to the verb 'request'

whereas *źu* [v.past] belongs to the verb 'melt'

Restoring ambiguity when a single form might belong to two distinct verbs

The preceding rules would make the change

źu [v.fut] ~ [v.past] ~ [v.pres]

> *źu* [v.invar]

but *źu* [v.fut] ~ [v.pres] belong to the verb 'request'

whereas *źu* [v.past] belongs to the verb 'melt'

So, we change

źu [v.invar]

Restoring ambiguity when a single form might belong to two distinct verbs

The preceding rules would make the change

źu [v.fut] ~ [v.past] ~ [v.pres]

> *źu* [v.invar]

but *źu* [v.fut] ~ [v.pres] belong to the verb 'request'

whereas *źu* [v.past] belongs to the verb 'melt'

So, we change

źu [v.invar]

> *źu* [v.fut.v.pres] ~ [v.past]

Other such cases

za [v.invar] 'itch'

versus *za* [v.pres] 'eat'

mchis [v.invar] 'be'

versus *mchis* [v.past] 'go'

How well does it work? Accuracy and Ambiguity

Mdzangs blun	Accuracy	Ambiguity
LexTagger	1.000	2.415
RuleTagger	0.988	1.403

Milarepa namthar	Accuracy	Ambiguity
LexTagger	1.000	2.441
RuleTagger	0.991	1.430

(on 11 Dec, 2013)

Thank you