Tibetan Corpus Linguistics:
present obstacles and future prospects

Nathan W. Hill and Edward Garrett
Tibetan in Digital Communication

Overview

1. Current analytic obstacles
Tibetan in Digital Communication

Overview

1. Current analytic obstacles
2. Software development desiderata
Overview

1. Current analytic obstacles
2. Software development desiderata
3. Future research questions
Tibetan in Digital Communication

Overview

1. Current analytic obstacles
2. Software development desiderata
3. Future research questions
4. Future collaborations
Tibetan in Digital Communication

Overview

1. Current analytic obstacles
2. Software development desiderata
3. Future research questions
4. Future collaborations
5. Possible projects
Current analytical obstacles

- Demonstrative uses of relator nouns
Current analytical obstacles

• Demonstrative uses of relator nouns
• Zero nominalization of verbs
Current analytical obstacles

- Demonstrative uses of relator nouns
- Zero nominalization of verbs
- Names
Current analytical obstacles

• Demonstrative uses of relator nouns
• Zero nominalization of verbs
• Names
• A few tricky words

rañ, re, etc.
Demonstrative uses of relator nouns

Demonstratives are sometimes used without an antecedent.

- *der tshogs-paḥi lha-mi*
  
The gods and men who had gathered there
Demonstrative uses of relator nouns

Demonstratives are sometimes used without an antecedent.

- *der tshogs-paḥi lha-mi*
  
The gods and men who had gathered there

We consequently treat some spatial words as demonstratives.
Demonstrative uses of relator nouns

Demonstratives are sometimes used without an antecedent.

- *der tshogs-paḥi lha-mi*
  
The gods and men who had gathered there

We consequently treat some spatial words as demonstratives.

- *phug-paḥi mdun-na mar log ḡgro-ẓiṅ ḡdug-pa-la*
  
While they were returning back down, passing in front of the cave.
Demonstrative uses of relator nouns

Demonstratives are sometimes used without an antecedent.

- *der tshogs-paḥi lha-mi*
  
The gods and men who had gathered there

We consequently treat some spatial words as demonstratives.

- *phug-paḥi mdun-na mar log ḷgro-ẑiṅ ḷdug-pa-la*
  
While they were returning back down, passing in front of the cave.

We consequently treat relator nouns, when they are not relating, as demonstratives.
Demonstrative uses of relator nouns

Demonstratives are sometimes used without an antecedent.

- **der tshogs-paḥi lha-mi**
  
The gods and men who had gathered there

We consequently treat some spatial words as demonstratives.

- **phug-paḥi mdun-na mar log ḥgro-zőṅ ḥdug-pa-la**
  
  While they were returning back down, passing in front of the cave.

We consequently treat relator nouns, when they are not relating, as demonstratives.

- **phi-na gos-daṅ / naṅ-na ḥtsho-ba phye-daṅ bcud ye med-paḥi stobs-kyis / lus keṅ-rus ltar son-ba /**
  
  Because neither clothes outside nor nourishment inside, whether food or drink, had he any, his body became like a skeleton.
Demonstrative uses of relator nouns

Demonstratives are sometimes used without an antecedent.

- *der tshogs-paḥi lha-mi*

  The gods and men who had gathered there

We consequently treat some spatial words as demonstratives.

- *phug-paḥi mdun-na mar log ḥgro-ziṅ ḥdug-pa-la*

  While they were returning back down, passing in front of the cave.

We consequently treat relator nouns, when they are not relating, as demonstratives.

- *phi-na gos-daṅ / naṅ-na ḥtsho-ba phye-daṅ bcud ye med-paḥi stobs-kyis / lus-keṅ-rus ltar soṅ-ba /*

  Because neither clothes outside nor nourishment inside, whether food or drink, had he any, his body became like a skeleton.

Maybe the spatial (temporal) uses of relator nouns and demonstratives should get a special tag?
Zero nominalization of verbs

In general Tibetan requires a suffix to change a verbal stem to a verbal noun, but not always, especially not in poetry.

- *bla-maḥi gsuṅ // ma-rig min-pa dbyiṅs-su dag /*

  The words of the guru, which are not ignorant, are as pure as space.
In general Tibetan requires a suffix to change a verbal stem to a verbal noun, but not always, especially not in poetry.

- **bla-mahi gsun // ma-rig min-pa dbyiṅs-su dag/**
  
  The words of the guru, which are not ignorant, are as pure as space.

- **ḥdir bzung gsan cig**
  
  'listen, O you who sit here!'
In general Tibetan requires a suffix to change a verbal stem to a verbal noun, but not always, especially not in poetry.

- **bla-maḥi gsuṅ // ma-rig min-pa dbyiṅs-su dag** /
  
  The words of the guru, which are not ignorant, are as pure as space.

- **ḥdir bzung gsan cig**
  
  'listen, O you who sit here!'

- **gzun-daṅ Ḍzin-paḥi sgrīb gniṅs bral**
  
  free from the two obscurations of 'taken' and 'taker'
Zero nominalization of verbs

In general Tibetan requires a suffix to change a verbal stem to a verbal noun, but not always, especially not in poetry.

- **bla-maḥi gsuṅ // ma-rig min-pa dbyiṅs-su dag /**
  
  The words of the guru, which are not ignorant, are as pure as space.

- **ḥdir bzung gsan cig**
  
  'listen, O you who sit here!'

- **gzuṅ-daṅ ḡdzin-paḥi sgrīb gñis bral**
  
  free from the two obscurations of 'taken' and 'taker'

- **ṅa-rgyal-daṅ ni ma-dad daṅ // don-du gñer-ba med ńid daṅ // phyi-rol-rnam-g.yeṅ-naṅ-bsdus daṅ // skyo-ba-ṅan-paḥi dri-ma yin //**
  
  Pride and lack of faith, lack of interest and being distracted outward, being withdrawn inward and dejection, (these) are flaws of listening.
Names

Names made up of more than one word make our word breaking inconsistent.

- *Chos kyi blo-gros*
  
  'Intelligence of dharma'
Names

Names made up of more than one word make our word breaking inconsistent.

- *Chos kyi blo-gros*
  'Intelligence of dharma'

- *Rgyal-po ḥi khab*
  'King's palace'
Names

Names made up of more than one word make our word breaking inconsistent.

- *Chos kyi blo-gros*
  
  'Intelligence of dharma'

- *Rgyal-po ḷi khab*
  
  'King's palace'

Ultimately this is solvable by tagging them as if they were not names and adding an additional workflow step of 'named entity recognition'.
A few tricky words

- raṅ
- [p.refl]

`raṅ gi bu-mo rnams sad-du btaṅ-ste` (he was) sent to wake his own daughters'

`khyed raṅ ‘you’, ña raṅ ‘I’, etc.`
A few tricky words

- raṅ

- [p.refl]

  raṅ gi bu-mo rnams sad-du btaṅ-ste 'he was) sent to wake his own daughters'

  khyed raṅ 'you', ňa raṅ 'I', etc.

- [p.pers]

  raṅ gis raṅ la lcag btab-nas 'they each hit one another'
A few tricky words

- rañ
- [p.refl]
  
  \textit{rañ gi bu-mo rnams sad-du btañ-ste} 'he was) sent to wake his own daughters'

  \textit{khyed rañ} ‘you’, \textit{ña rañ} ‘I’, etc.

- [p.pers]

  \textit{rañ gis rañ la lcag btab-nas} 'they each hit one another'

- [d.det]

  \textit{khyod las rem-po rañ źig} 'one more assiduous even than thou'
A few tricky words

- raṅ
  - [p.refl]
    
    **raṅ gi bu-mo rnams sad-du btar-ste** 'he was) sent to wake his own daughters'
    
    **khyed raṅ 'you', ni raṅ 'I', etc.**
  
  - [p.pers]
    
    **raṅ gis raṅ la lcag btab-nas** 'they each hit one another'
    
  - [d.det]
    
    **khyod las rem-po raṅ žig** 'one more assiduous even than thou'

  - [???]
    
    **la-la na-re Mar-pa raṅ smyos-nas** 'some said Marpa (himself) had gone (quite) crazy'.

A few tricky words

• \textit{re}
• \texttt{[p.indef]}

\textit{re rāñ} 'each' (we use this tag inconsistently)
A few tricky words

- **re**
  - [p.indef]
    
    _re raiñ_ 'each' (we use this tag inconsistently)
  
  - [d.det]

    _yum-gyis nas khal re yod-pahi chañ tshod chen-po gsum btsos-pa-la_
    'mother cooked up each load of barley that she had into three great measures of beer'
A few tricky words

• *re*

  • [p.indef]
    
    *re rāñ* 'each' (we use this tag inconsistently)

  • [d.det]

    *yum-gyis nas khal re yod-pahi chañ tshod chen-po gsum btsos-pa-la*
    
    'mother cooked up each load of barley that she had into three great measures of beer'

• [n.count]

  *re ŋig* 'one time', *re* 'hope'
A few tricky words

- re
- [p.indef]
  re rañ 'each' (we use this tag inconsistently)
- [d.det]
  yum-gyis nas khal re yod-pahi chañ tshod chen-po gsum btsos-pa-la
  'mother cooked up each load of barley that she had into three great measures of beer'
- [n.count]
  re žig 'one time', re 'hope'
- [num.card]
  žo re Žo do 'one or two ounces'
Software development desiderata

- Automation of recompiling and retagging
Software development desiderata

- Automation of recompiling and retagging
- Incorporation of Pablo's consistency tester
Software development desiderata

• Automation of recompiling and retagging
• Incorporation of Pablo's consistency tester
• Post hoc rule based consistency checker
Software development desiderata

- Automation of recompiling and retagging
- Incorporation of Pablo's consistency tester
- Post hoc rule based consistency checker
- Seamless web interface
Post hoc rule based consistency checker

- Negation and verb stems

  \textit{mi} [neg] \textit{dgaĥ-ba} [n.v.fut.n.v.pres] 'not happy'

  \textit{mi} [n.count] \textit{dgaĥ-ba} [n.v.invar] 'happy person'
Post hoc rule based consistency checker

- Negation and verb stems

  \textit{mi} [neg] \textit{dgah-ba} [n.v.fut.n.v.pres] 'not happy'

  \textit{mi} [n.count] \textit{dgah-ba} [n.v.invar] 'happy person'

  *\textit{mi} [n.count] \textit{dgah-ba} [n.v.fut.n.v.pres]

  *\textit{mi} [neg] \textit{dgah-ba} [n.v.invar]
Post hoc rule based consistency checker

- Negation and verb stems

  *mi* [neg] *dgaḥ-ba* [n.v.fut.n.v.pres] 'not happy'

  *mi* [n.count] *dgaḥ-ba* [n.v.invar] 'happy person'

  *mi* [n.count] *dgaḥ-ba* [n.v.fut.n.v.pres]

  *mi* [neg] *dgaḥ-ba* [n.v.invar]

- If *mi* is [n.count] then change *dgaḥ-ba* [n.v.fut.n.v.pres] to *dgaḥ-ba* [n.v.invar]
Post hoc rule based consistency checker

- Negation and verb stems

  \( mi \) [neg] \textit{dga\-ba} [n.v.fut.n.v.pres] 'not happy'

  \( mi \) [n.count] \textit{dga\-ba} [n.v.invar] 'happy person’

  *\( mi \) [n.count] \textit{dga\-ba} [n.v.fut.n.v.pres]

  *\( mi \) [neg] \textit{dga\-ba} [n.v.invar]

- If \( mi \) is [n.count] then change \textit{dga\-ba} [n.v.fut.n.v.pres] to \textit{dga\-ba} [n.v.invar]

- If \( mi \) is [neg] then change \textit{dga\-ba} [n.v.invar] to \textit{dga\-ba} [n.v.fut.n.v.pres]
Post hoc rule based consistency checker

\( de \) [adv.proclausal] \( ūi \) [case.gen] \( tshe \) [n.re] 'at that time'

\( de \) [d.dem] \( ūi \) [case.gen] \( tshe \) [n.count] 'his life'
Post hoc rule based consistency checker

*de [adv.proclausal] ʰi [case.gen] tshe [n.rel] 'at that time'


*de [d.dem] ʰi [case.gen] tshe [n.rel]

*de [adv.proclausal] ʰi [case.gen] tshe [n.count]
Post hoc rule based consistency checker

*de* [adv.proclausal] *hi* [case.gen] *tshe* [n.rel] 'at that time'
*de* [d.dem] *hi* [case.gen] *tshe* [n.count] 'his life'
*de* [adv.proclausal] *hi* [case.gen] *tshe* [n.rel]
*de* [adv.proclausal] *hi* [case.gen] *tshe* [n.count]

- If *tshe* is [n.count] then change *de* [adv.proclausal] to *de* [d.dem]
Research questions we can address with our Corpus

1. Conflicts between *da-drag* rules and syntactic cues to verb stem disambiguation.
1. Conflicts between *da-drag* rules and syntactic cues to verb stem disambiguation.

"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."
Research questions we can address with our Corpus

1. Conflicts between da-drag rules and syntactic cues to verb stem disambiguation.

   ku-šu 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].
Research questions we can address with our Corpus

1. Conflicts between da-drag rules and syntactic cues to verb stem disambiguation.

   ku-śu ḡdi ni ḡbras-bu las skyes-pa ma lags te | chab-mig cig gi nari nas rñed-pa s slan-cad ni bdag gis mi rñed de | mi ḡbyor [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 ḡbyor is tagged [v.pres].)
Research questions we can address with our Corpus

1. Conflicts between *da-drag* rules and syntactic cues to verb stem disambiguation.

2. See whether any converbs imply things about what verb stems can be coordinated (e.g. *V-pa-daň* … *V-pa-daň*, do they have to be the same stem?)
Research questions we can address with our Corpus

1. Conflicts between *da-drag* rules and syntactic cues to verb stem disambiguation.

2. See whether any converbs imply things about what verb stems can be coordinated (e.g. *V-pa-dañ* … *V-pa-dañ*, do they have to be the same stem?)

3. Can we identify different uses of *śad* by counting how many words occur between them?
Research questions we can address with our Corpus

1. Conflicts between *da-drag* rules and syntactic cues to verb stem disambiguation.

2. See whether any converbs imply things about what verb stems can be coordinated (e.g. *V-pa-dañ* … *V-pa-dañ*, do they have to be the same stem?)

3. Can we identify different uses of *śad* by counting how many words occur between them?

Research questions we can address with our Corpus

1. Conflicts between *da-drag* rules and syntactic cues to verb stem disambiguation.

2. See whether any converbs imply things about what verb stems can be coordinated (e.g. *V-pa-dañ* ... *V-pa-dañ*, do they have to be the same stem?)

3. Can we identify different uses of *śad* by counting how many words occur between them?


5. Functions of *la* [cv.all] other than to coordinate imperatives.
Future Collaborations

- TBRC
- Berkeley
- CASS
- You!
Possible Projects

• Tibet Daily Corpus (e.g. for keywords in current affairs)
Possible Projects

- Tibet Daily Corpus (e.g. for keywords in current affairs)
- Continuous crawling of Tibetan blogs (e.g. for new terminology)
Possible Projects

- Tibet Daily Corpus (e.g. for keywords in current affairs)
- Continuous crawling of Tibetan blogs (e.g. for new terminology)
- Yours?
Thank you