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

Tibetan in Digital Communication

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

  raṅ, 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 ḥgro-źiṅ ḥdug-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 ḥgro-źiṅ 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.

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-źiṅ 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.

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-źiṅ 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.

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 **bźugs** gsan cig*

  'listen, O you who sit here!'

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 **bźugs** gsan cig*

  'listen, O you who sit here!'

- ***gzuṅ**-daṅ ḥdzin-paḥi sgrib gñis 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 **bźugs** gsan cig*

  'listen, O you who sit here!'

- ***gzuṅ**-daṅ ḥdzin-paḥi sgrib gñis bral*

  free from the two obscurations of 'taken' and 'taker'

- *ṅa-rgyal-daṅ ni ma-dad daṅ // don-du gñer-ba **med** ñid daṅ // phyi-rol-rnam-g.yeṅ-naṅ-bsdus daṅ // skyo-ba-ñan-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 ḥi 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’***, ṅa* ***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’***, ṅa **raṅ*** **‘I’**, etc.
  - [p.pers]

    ***raṅ** gis **raṅ** la lcag btab-nas* 'they each hit one another'

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’**, *ṅa* ***raṅ*** **‘I’**, etc.
  - [p.pers]

    raṅ *gis* ***raṅ*** *la lcag btab-nas* 'they each hit one another'
  - [d.det]

    khyod las rem-po ***raṅ*** *źig* 'one more assiduous even than thou'

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’***, ṅa* ***raṅ*** **‘I’**, etc.
  - [p.pers]

    ***raṅ** gis **raṅ** la lcag btab-nas* 'they each hit one another'
  - [d.det]

    khyod las rem-po ***raṅ*** *źig* '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

- *re*
  - [p.indef]

    re *raṅ* 'each' (we use this tag inconsistently)

A few tricky words

- *re*
  - [p.indef]

    re *raṅ* 'each' (we use this tag inconsistently)

  - [d.det]

    yum-gyis nas khal ***re*** *yod-paḥi chaṅ 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 *raṅ* 'each' (we use this tag inconsistently)

  - [d.det]

    yum-gyis nas khal ***re*** *yod-paḥi chaṅ 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 *źig* 'one time', ***re*** 'hope'

A few tricky words

- *re*
  - [p.indef]

    ***re** rań* 'each' (we use this tag inconsistently)

  - [d.det]

    *yum-gyis nas khal **re** yod-paḥi chań 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** źig* 'one time', ***re 'hope'

  - [num.card]

    *źo **re** źo 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

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

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

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]

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] *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]
- If *mi* is [neg] then change *dgaḥ-ba* [n.v.invar] to *dgaḥ-ba* [n.v.fut.n.v.pres]

Post hoc rule based consistency checker

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

de [d.dem] *ḥi* [case.gen] *tshe* [n.count] 'his life'

Post hoc rule based consistency checker

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

de [d.dem] *ḥi* [case.gen] *tshe* [n.count] 'his life’

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

*de [adv.proclausal] *ḥi* [case.gen] *tshe* [n.count]

Post hoc rule based consistency checker

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

de [d.dem] *ḥi* [case.gen] *tshe* [n.count] 'his life’

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

*de [adv.proclausal] *ḥi* [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.

Research questions we can address with our Corpus

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

ku-śu ḥdi ni ḥbras-bu las skyes-pa ma lags te ǀ chab-mig cig gi naṅ nas rñed-pa s slan-cad ni bdag gis mi rñed de ǀ 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."

Research questions we can address with our Corpus

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

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].

Research questions we can address with our Corpus

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

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].)

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-daṅ* … V-*pa-daṅ*, 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-daṅ* … V-*pa-daṅ*, 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-daṅ* … V-*pa-daṅ*, do they have to be the same stem?)

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

4. How do conditional clauses work (cf. Greek).

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-daṅ* … V-*pa-daṅ*, do they have to be the same stem?)

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

4. How do conditional clauses work (cf. Greek).

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