Tibetan Corpus Linguistics: our progress so far

Nathan W. Hill and Edward Garrett

(SOAS, University of London)

Introducing Corpus Linguistics

Maslow's hierarchy of Corpus Linguistic needs

Introducing Corpus Linguistics

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode

Introducing Corpus Linguistics

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode
2. Some digital texts are available

Introducing Corpus Linguistics

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode
2. Some digital texts are available
3. Segmenter: Words are divided (orthographically or with software)

Introducing Corpus Linguistics

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode
2. Some digital texts are available
3. Segmenter: Words are divided
4. Tagger: Part of speech of each word is identifiable

Introducing Corpus Linguistics

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode
2. Some digital texts are available
3. Segmenter: Words are divided
4. Tagger: Part of speech of each word is identifiable
    1. E.g. ‘sit on a chair’ [noun] versus ‘chair a meeting’ [verb]

Introducing Corpus Linguistics

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode
2. Some digital texts are available
3. Segmenter: Words are divided
4. Tagger: Part of speech of each word is identifiable
5. Lemmatizer: Different forms of a word are associated with each other

Introducing Corpus Linguistics

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode
2. Some digital texts are available
3. Segmenter: Words are divided
4. Tagger: Part of speech of each word is identifiable
5. Lemmatizer: Different forms of a word are associated with each other
    1. ‘sing’, ‘sang’, ‘sung’, ‘singing’, ‘sings’ all associated with [sing]

Introducing Corpus Linguistics

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode
2. Some digital texts are available
3. Segmenter: Words are divided
4. Tagger: Part of speech of each word is identifiable
5. Lemmatizer: Different forms of a word are associated with each other
6. Parser: Higher order syntactic analysis
    1. E.g. noun phrase detection, verbal rection, etc.

E-resources for English

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode
2. Some digital texts are available
3. Segmenter: Words are divided
4. Tagger: Part of speech of each word is identifiable
5. Lemmatizer: Different forms of a word are associated with each other
6. Parser: Higher order syntactic analysis

We have all of them!

E-resources for Tibetan

Maslow's hierarchy of Corpus Linguistic needs

E-resources for Tibetan

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode

E-resources for Tibetan

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode ✓

E-resources for Tibetan

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode ✓
2. Some digital texts are available

E-resources for Tibetan

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode ✓
2. Some digital texts are available ✓

E-resources for Tibetan

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode ✓
2. Some digital texts are available ✓
3. Segmenter: Words are divided
4. Tagger: Part of speech of each word is identifiable
5. Lemmatizer: Different forms of a word are associated with each other
6. Parser: Higher order syntactic analysis

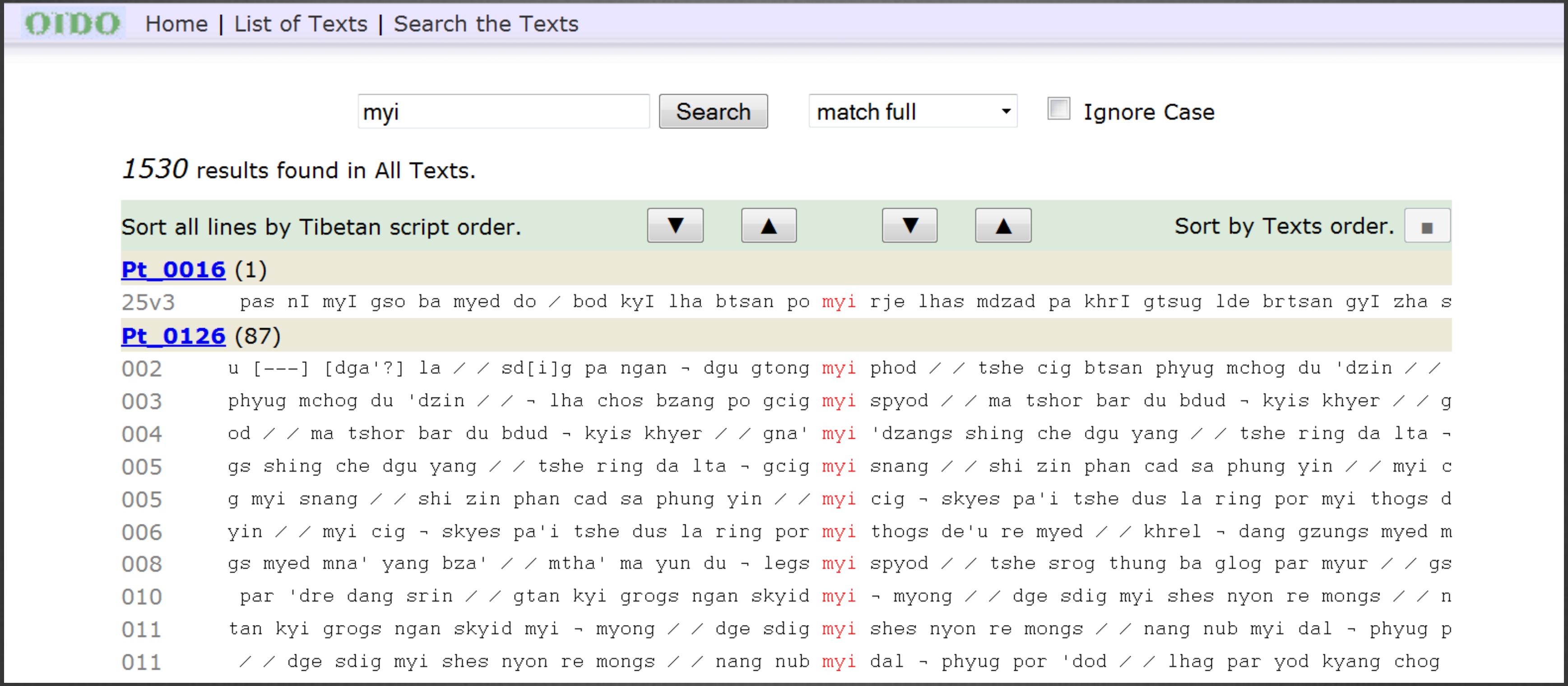
E-resources for Tibetan

Maslow's hierarchy of Corpus Linguistic needs

1. Script is in Unicode ✓
2. Some digital texts are available ✓
3. Segmenter: Words are divided
4. Tagger: Part of speech of each word is identifiable
5. Lemmatizer: Different forms of a word are associated with each other
6. Parser: Higher order syntactic analysis

Our focus

Tibetan e-resources:
Old Tibetan Documents Online (OTDO)



myi

1530 results found in All Texts.

Sort all lines by Tibetan script order.

Sort by Texts order.

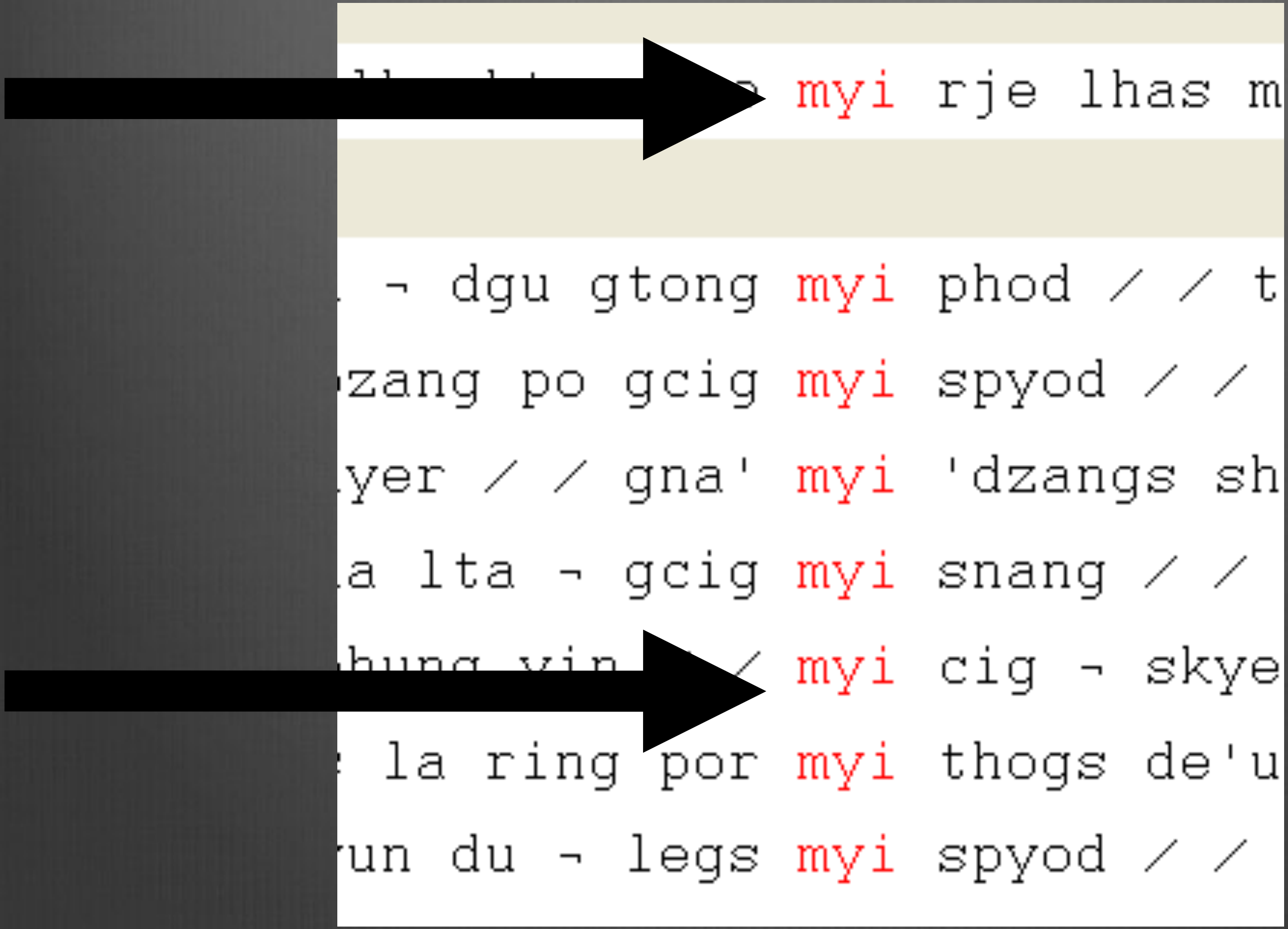
Pt_0016 (1)

```
25v3     pas nI myI gso ba myed do ⁄ bod kyI lha btsan po myi rje lhas mdzad pa khrI gtsug lde brtsan gyI zha s
```

Pt_0126 (87)

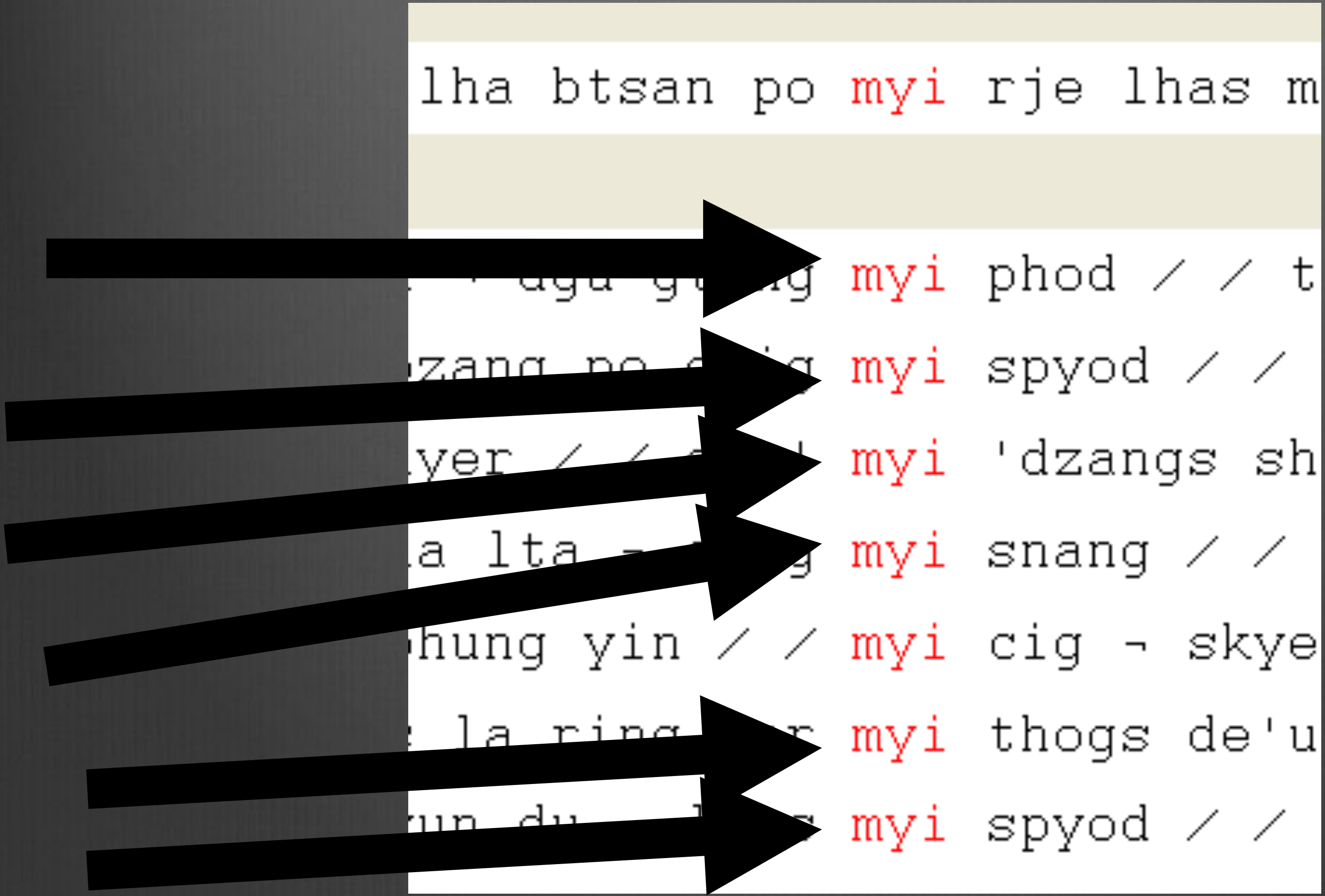
```
002     u [---] [dga'?] la ⁄ ⁄ sd[i]g pa ngan ¬ dgu gtong myi phod ⁄ ⁄ tshe cig btsan phyug mchog du 'dzin ⁄ ⁄
003     phyug mchog du 'dzin ⁄ ⁄ ¬ lha chos bzang po gcig myi spyod ⁄ ⁄ ma tshor bar du bdud ¬ kyis khyer ⁄ ⁄ g
004     od ⁄ ⁄ ma tshor bar du bdud ¬ kyis khyer ⁄ ⁄ gna' myi 'dzangs shing che dgu yang ⁄ ⁄ tshe ring da lta ¬
005     gs shing che dgu yang ⁄ ⁄ tshe ring da lta ¬ gcig myi snang ⁄ ⁄ shi zin phan cad sa phung yin ⁄ ⁄ myi c
005     g myi snang ⁄ ⁄ shi zin phan cad sa phung yin ⁄ ⁄ myi cig ¬ skyes pa'i tshe dus la ring por myi thogs d
006     yin ⁄ ⁄ myi cig ¬ skyes pa'i tshe dus la ring por myi thogs de'u re myed ⁄ ⁄ khrel ¬ dang gzungs myed m
008     gs myed mna' yang bza' ⁄ ⁄ mtha' ma yun du ¬ legs myi spyod ⁄ ⁄ tshe srog thung ba glog par myur ⁄ ⁄ gs
010      par 'dre dang srin ⁄ ⁄ gtan kyi grogs ngan skyid myi ¬ myong ⁄ ⁄ dge sdig myi shes nyon re mongs ⁄ ⁄ n
011     tan kyi grogs ngan skyid myi ¬ myong ⁄ ⁄ dge sdig myi shes nyon re mongs ⁄ ⁄ nang nub myi dal ¬ phyug p
011      ⁄ ⁄ dge sdig myi shes nyon re mongs ⁄ ⁄ nang nub myi dal ¬ phyug por 'dod ⁄ ⁄ lhag par yod kyang chog
```

མྱི་ *myi* 'person'



```
                  myi rje lhas m
. ¬ dgu gtong myi phod / / t
zang po gcig myi spyod / /
yer / / gna' myi 'dzangs sh
a lta ¬ gcig myi snang / /
hung yin  / myi cig ¬ skye
 la ring por myi thogs de'u
un du ¬ legs myi spyod / /
```

མྱི་ *myi* 'not'



lha btsan po myi rje lhas m

: ' dgu g ng myi phod / / t

zang po c g myi spyod / /

yer / / ' myi 'dzangs sh

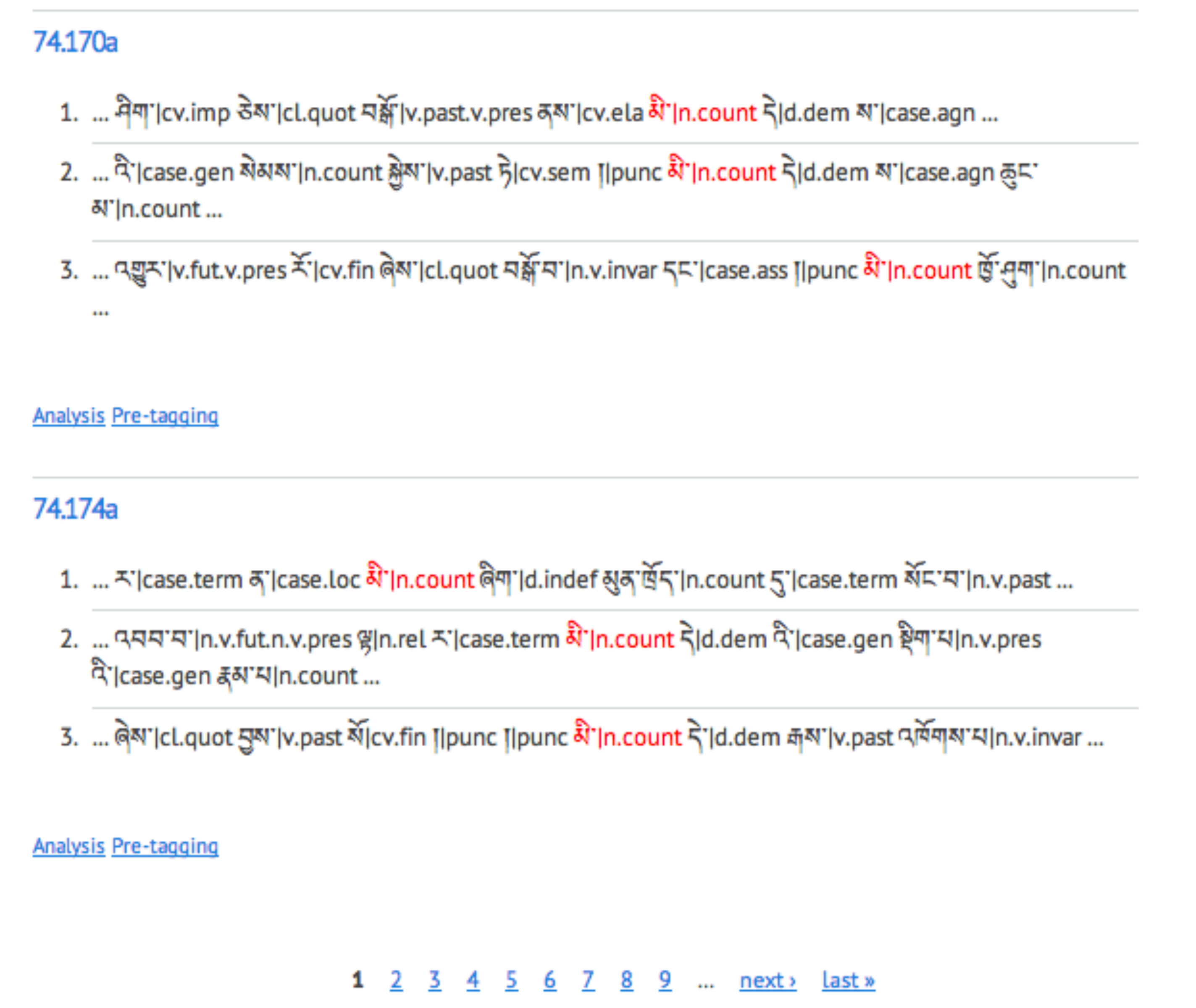
a lta - g myi snang / /

hung yin / / myi cig ¬ skye

: la ring r myi thogs de'u

un du g myi spyod / /

A second try with མི་ *mi* 'person'



74.170a

1. ... ཤིག་|cv.imp ཅེས་|cl.quot བསྒོ|v.past.v.pres ནས་|cv.ela མི་|n.count དེ|d.dem ས་|case.agn ...
2. ... འི་|case.gen སེམས་|n.count སྐྱེས་|v.past ཏེ|cv.sem །|punc མི་|n.count དེ|d.dem ས་|case.agn ཆུང་མ་|n.count ...
3. ... འགྱུར་|v.fut.v.pres རོ་|cv.fin ཞེས་|cl.quot བསྒོ་བ་|n.v.invar དང་|case.ass །|punc མི་|n.count ཁྱོ་ཤུག་|n.count ...

Analysis Pre-tagging

74.174a

1. ... ར་|case.term ན་|case.loc མི་|n.count ཞིག་|d.indef སྨན་ཁྱོད་|n.count ཏུ་|case.term སོང་བ་|n.v.past ...
2. ... འབབ་བ་|n.v.fut.n.v.pres ལྷ|n.rel ར་|case.term མི་|n.count དེ|d.dem འི་|case.gen སྡིག་པ|n.v.pres འི་|case.gen རྣམ་པ|n.count ...
3. ... ཞེས་|cl.quot བྱས་|v.past སོ|cv.fin །|punc །|punc མི་|n.count དེ་|d.dem རྐུས་|v.past འཁོགས་པ|n.v.invar ...

Analysis Pre-tagging

1 2 3 4 5 6 7 8 9 ... next › last »

Tibetan in Digital Communication

Goals

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

Our Corpora

Classical

Mdzaṅs-blun
9th century canonical narrative trans. from Chinese (55,059+ words)

Bu ston chos ḥbyuṅ
13th century history, mostly quotes from earlier sources (89,129)

Mi-la ras-paḥi rnam thar
15th century biography (41,864+ words)

Mar-paḥi rnam thar
15th century biography (39,969+ words)

Pavel

39,011 words of various texts

Balk

85,143 catalog of Berlin Tibetica

POS tag set

The POS tag set will not be discussed much today.

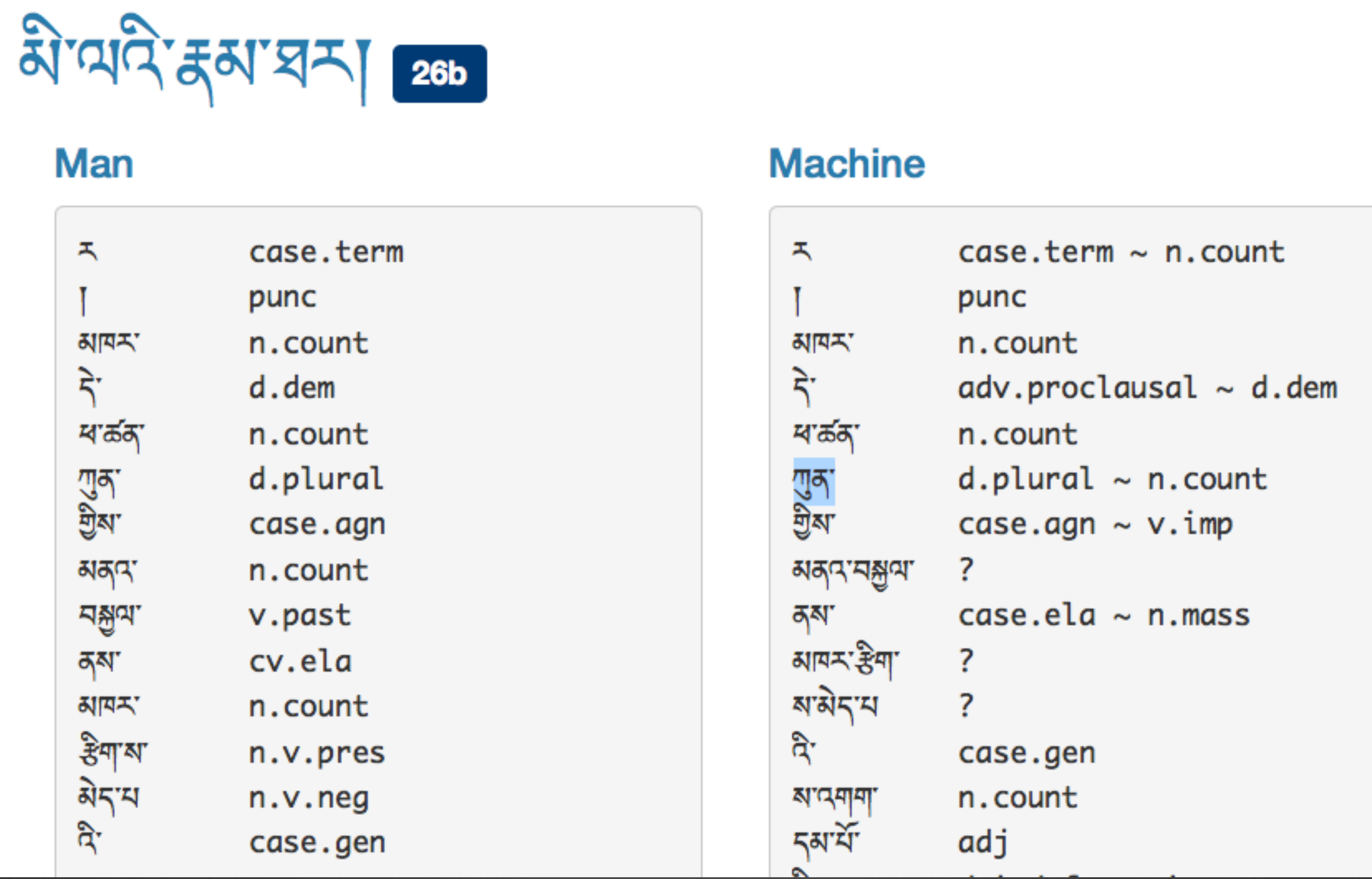
Garrett, Edward and Hill, Nathan W. and Kilgarriff, Adam and Vadlapudi, Ravikiran and Zadoks, Abel (2015). "The contribution of corpus linguistics to lexicography and the future of Tibetan dictionaries." *Revue d'Etudes Tibétaines* 32: 51-86.

Word breaking

Our Achilles heel

0.92397 accurate (15 April, 2015)

Workflow:



མི་ལའི་རྣམ་ཐར། 26b

Man

ར	case.term
།	punc
མཁར་	n.count
དེ་	d.dem
ཕ་ཚན་	n.count
ཀུན་	d.plural
གྱིས་	case.agn
མནའ་	n.count
བསྐྱལ་	v.past
ནས་	cv.ela
མཁར་	n.count
རྩིག་ས་	n.v.pres
མེད་པ	n.v.neg
འི་	case.gen

Machine

ར	case.term ~ n.count
།	punc
མཁར་	n.count
དེ་	adv.proclausal ~ d.dem
ཕ་ཚན་	n.count
ཀུན་	d.plural ~ n.count
གྱིས་	case.agn ~ v.imp
མནའ་བསྐྱལ་	?
ནས་	case.ela ~ n.mass
མཁར་རྩིག་	?
ས་མེད་པ	?
འི་	case.gen
ས་འགག་	n.count
དམ་པོ་	adj

Workflow:
(1) Look-up of possible analyses

Word	Transliteration	Part-of-speech tag
རྒྱལ་པོ་	*rgyal-po*	n.count
དེ་	*de*	d.dem ∼ cv.sem
ལ་	*la*	case.all ∼ n.count
བཙུན་མོ་	*btsun-mo*	n.count
ལྔ་	*lṅa*	num.card
བརྒྱ་	*brgya*	num.card
ཡོད་	*yod*	v.invar
ཀྱང་	*kyaṅ*	cl.focus
།		punc

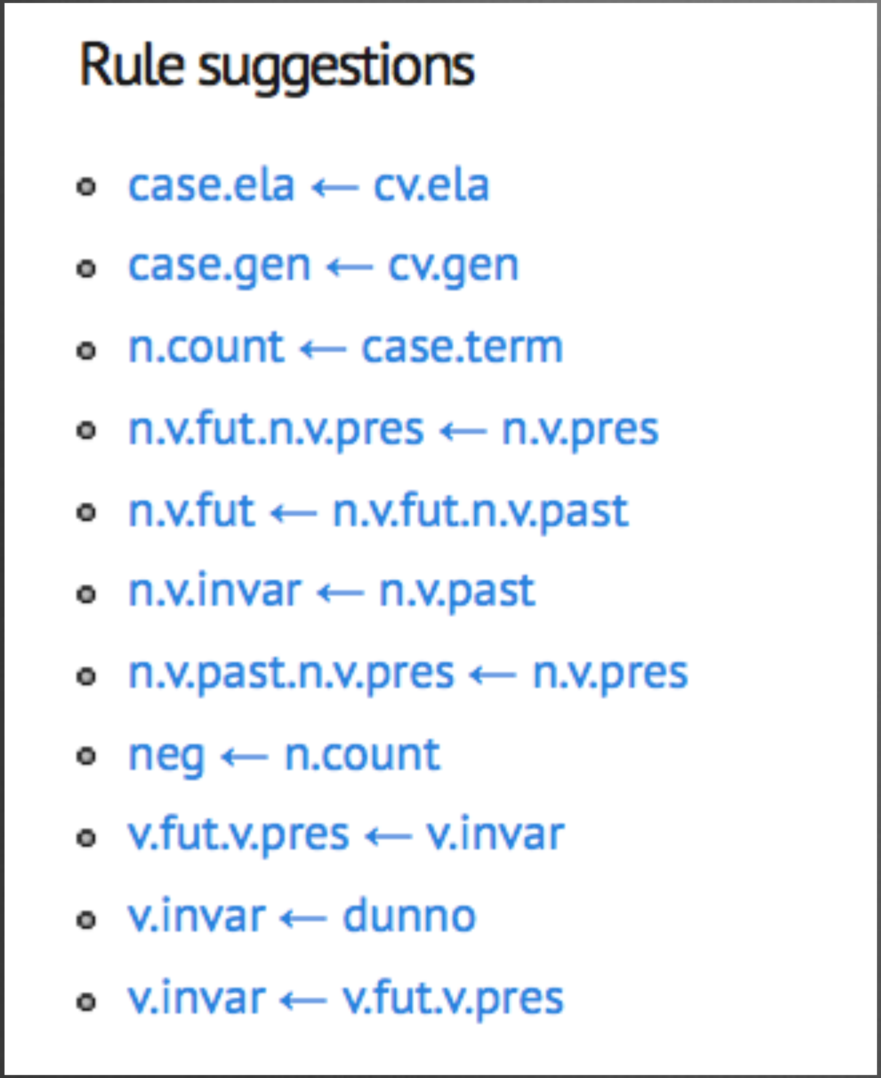
Workflow:
(2) Pre-tagging

Word	Transliteration	Part-of-speech tag
རྒྱལ་པོ་	*rgyal-po*	n.count
དེ་	*de*	d.dem
ལ་	*la*	case.all ∼ n.count
བཙུན་མོ་	*btsun-mo*	n.count
ལྔ་	*lṅa*	num.card
བརྒྱ་	*brgya*	num.card
ཡོད་	*yod*	v.invar
ཀྱང་	*kyaṅ*	cl.focus
།		punc

Workflow:
(3) Hand-tagging

Word	Transliteration	Part-of-speech tag
རྒྱལ་པོ་	*rgyal-po*	n.count
དེ་	*de*	d.dem
ལ་	*la*	case.all
བཙུན་མོ་	*btsun-mo*	n.count
ལྔ་	*lṅa*	num.card
བརྒྱ་	*brgya*	num.card
ཡོད་	*yod*	v.invar
ཀྱང་	*kyaṅ*	cl.focus
।		punc

Workflow:
(4) Rule suggestions



Rule suggestions

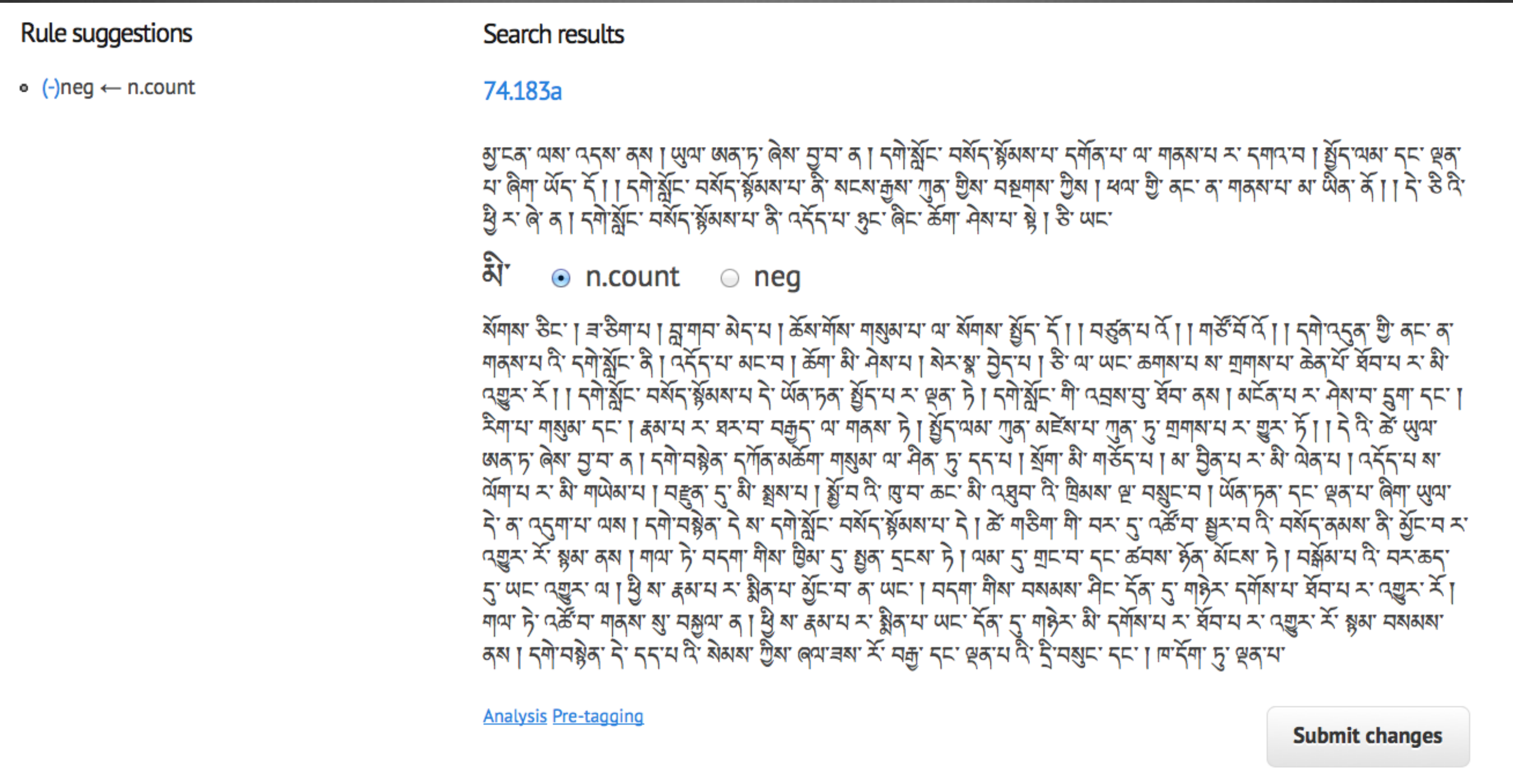
- case.ela ← cv.ela
- case.gen ← cv.gen
- n.count ← case.term
- n.v.fut.n.v.pres ← n.v.pres
- n.v.fut ← n.v.fut.n.v.past
- n.v.invar ← n.v.past
- n.v.past.n.v.pres ← n.v.pres
- neg ← n.count
- v.fut.v.pres ← v.invar
- v.invar ← dunno
- v.invar ← v.fut.v.pres

Screen shot of rule suggestions

(9 November 2013)

Workflow:
(4) Rule suggestions

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



Rule suggestions

- (-)neg ← n.count

Search results

74.183a

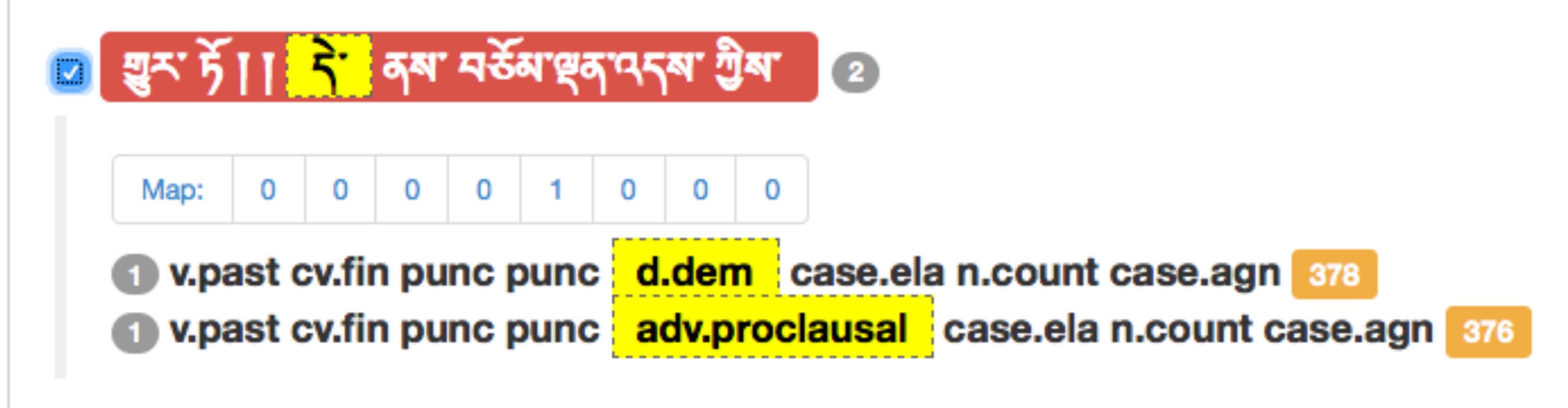
མྱ་ངན་ ལས་ འདས་ ནས ། ཡུལ་ ཨན་ཏ་ ཞེས་ བྱ་བ་ ན ། དགེ་སློང་ བསོད་སྙོམས་པ་ དགོན་པ་ ལ་ གནས་པ་ ར་ དཀའ་བ ། སྤྱོད་ལམ་ དང་ ལྡན་ པ་ ཞིག་ ཡོད་ དོ ། ། དགེ་སློང་ བསོད་སྙོམས་པ་ ནི་ སངས་རྒྱས་ ཀུན་ གྱིས་ བསྔགས་ ཀྱིས ། ཕལ་ གྱི་ ནང་ ན་ གནས་པ་ མ་ ཡིན་ ནོ ། ། དེ་ ཅི་ འི་ ཕྱི་ ར་ ཞེ་ ན ། དགེ་སློང་ བསོད་སྙོམས་པ་ ནི་ འདོད་པ་ ཆུང་ ཞིང་ ཆོག་ ཤེས་པ་ སྟེ ། ཅི་ ཡང་

མི་ ● n.count ○ neg

སོགས་ ཅིང་ ། ཟ་ཅིག་པ ། བླ་གབ་ མེད་པ ། ཆོས་གོས་ གསུམ་པ་ ལ་ སོགས་ སྦྱོང་ ངོ ། ། བཙུན་པ་ འོ ། ། གཙོ་བོ་ འོ ། ། དགེ་འདུན་ གྱི་ ནང་ ན་ གནས་པ་ འི་ དགེ་སློང་ ནི ། འདོད་པ་ མང་བ ། ཆོག་ མི་ ཤེས་པ ། སེར་སྣ་ བྱེད་པ ། ཅི་ ལ་ ཡང་ ཆགས་པ་ ས་ གྲགས་པ་ ཆེན་པོ་ ཐོབ་པ་ ར་ མི་ འགྱུར་ རོ ། ། དགེ་སློང་ བསོད་སྙོམས་པ་ དེ་ ཡོན་ཏན་ སྦྱོང་པ་ ར་ ལྡན་ ཏེ ། དགེ་སློང་ གི་ འབྲས་བུ་ ཐོབ་ ནས ། མངོན་པ་ ར་ ཤེས་པ་ དྲུག་ དང་ ། རིག་པ་ གསུམ་ དང་ ། རྣམ་པ་ ར་ ཐར་བ་ བརྒྱད་ ལ་ གནས་ ཏེ ། སྤྱོད་ལམ་ ཀུན་ མཛེས་པ་ ཀུན་ ཏུ་ གྲགས་པ་ ར་ གྱུར་ ཏོ ། ། དེ་ འི་ ཚེ་ ཡུལ་ ཨན་ཏ་ ཞེས་ བྱ་བ་ ན ། དགེ་བསྙེན་ དཀོན་མཆོག་ གསུམ་ ལ་ ཤིན་ ཏུ་ དད་པ ། སྲོག་ མི་ གཅོད་པ ། མ་ བྱིན་པ་ ར་ མི་ ལེན་པ ། འདོད་པ་ ས་ ལོག་པ་ ར་ མི་ གཡེམ་པ ། བརྫུན་ དུ་ མི་ སྨྲ་པ ། སྨྱོ་བ་ འི་ ཁུ་བ་ ཆང་ མི་ འཐུབ་ འི་ ཁྲིམས་ ལྔ་ བསྲུང་བ ། ཡོན་ཏན་ དང་ ལྡན་པ་ ཞིག་ ཡུལ་ དེ་ ན་ འདུག་པ་ ལས ། དགེ་བསྙེན་ དེ་ ས་ དགེ་སློང་ བསོད་སྙོམས་པ་ དེ ། ཚེ་ གཅིག་ གི་ བར་ དུ་ འཚོ་བ་ སྦྱར་བ་ འི་ བསོད་ནམས་ ནི་ མྱོང་བ་ ར་ འགྱུར་ རོ་ སྙམ་ ནས ། གལ་ ཏེ་ བདག་ གིས་ ཁྱིམ་ དུ་ སྤྱན་ དྲངས་ ཏེ ། ལམ་ དུ་ གྲང་བ་ དང་ ཚབས་ ཉེན་ མོངས་ ཏེ ། བསྒོམ་པ་ འི་ བར་ཆད་ དུ་ ཡང་ འགྱུར་ ལ ། ཕྱི་ ས་ རྣམ་པ་ ར་ སྨིན་པ་ མྱོང་བ་ ན་ ཡང་ ། བདག་ གིས་ བསམས་ ཤིང་ དོན་ དུ་ གཉེར་ དགོས་པ་ ཐོབ་པ་ ར་ འགྱུར་ རོ ། གལ་ ཏེ་ འཚོ་བ་ གནས་ སུ་ བསྐྱལ་ ན ། ཕྱི་ ས་ རྣམ་པ་ ར་ སྨིན་པ་ ཡང་ དོན་ དུ་ གཉེར་ མི་ དགོས་པ་ ར་ ཐོབ་པ་ ར་ འགྱུར་ རོ་ སྙམ་ བསམས་ ནས ། དགེ་བསྙེན་ དེ་ དད་པ་ འི་ སེམས་ ཀྱིས་ ཞལ་ཟས་ རོ་ བརྒྱ་ དང་ ལྡན་པ་ འི་ དྲི་བསུང་ དང་ ། ཁ་དོག་ ཏུ་ ལྡན་པ

Workflow:
(5) Checking consistency

Using a programme provided by Pablo Faria of UNICAMP.



☑ སྦྱར་ ཏོ། ། དེ་ ནས་ བཅོམ་ལྡན་འདས་ ཀྱིས་ 2

Map:	0	0	0	0	1	0	0	0

1 v.past cv.fin punc punc d.dem case.ela n.count case.agn 378

1 v.past cv.fin punc punc adv.proclausal case.ela n.count case.agn 376

Does *de nas* mean ‘from him’ or ‘then’?

Disambiguating *mi* as negation or a noun

Isolating *mi* [n.count] after the genitive

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

Disambiguating *mi* as negation or a noun

Isolating *mi* [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* as negation or a noun

Isolating *mi* [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* as negation or a noun

Isolating *mi* [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* 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* as a [n.count].

Disambiguating *mi* as negation or a noun

Isolating *mi* [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* 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* as a [n.count].

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

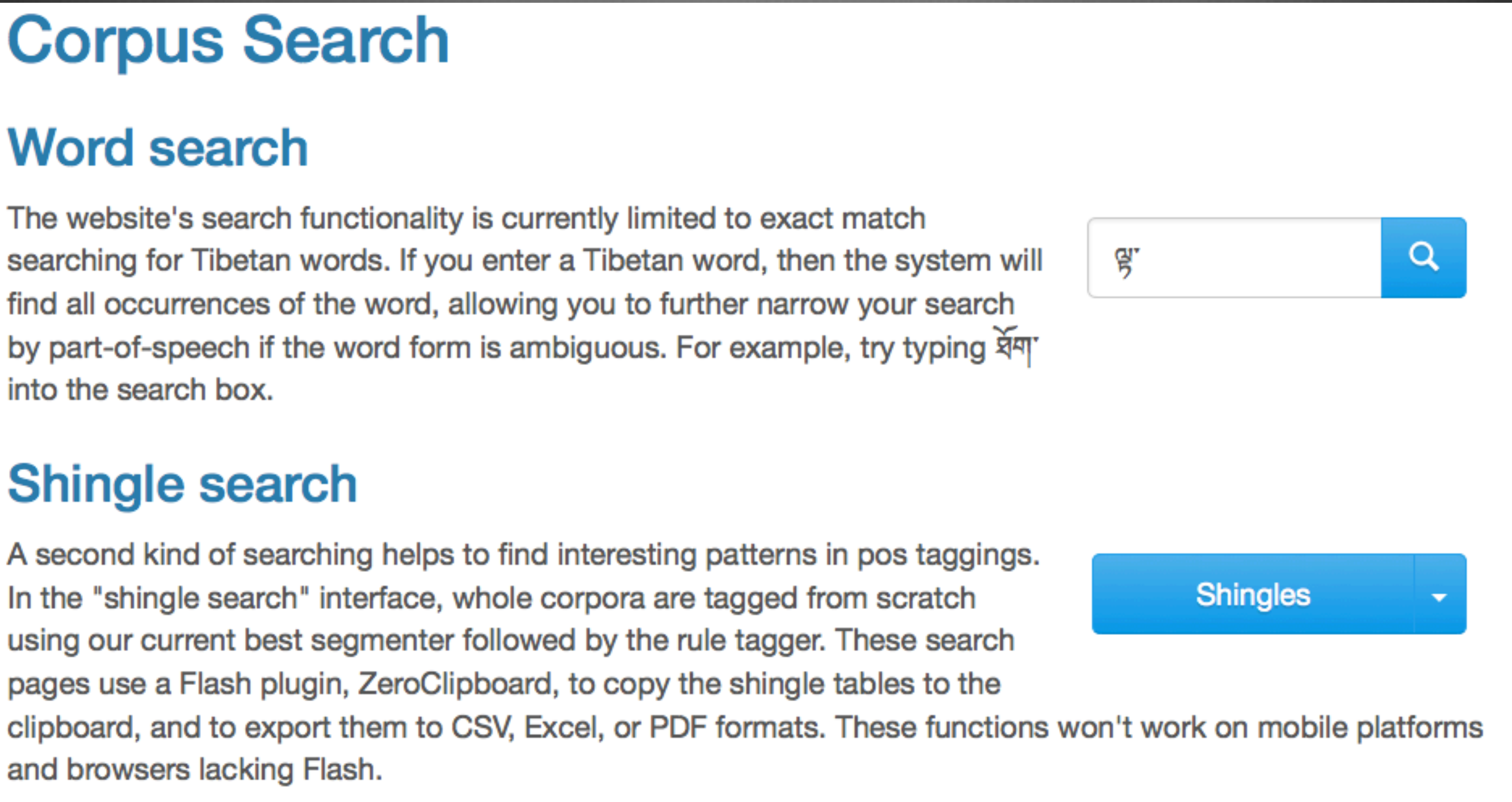
REPLACE: $1|[n.count]

The rule based tagger

For more about the rule based tagger—

Garrett, Edward and Hill, Nathan W. and Zadoks, Abel (2014) 'A Rule-based Part-of-speech Tagger for Classical Tibetan.' *Himalayan Linguistics*, 13 (1). pp. 9-57.

Search



Corpus Search

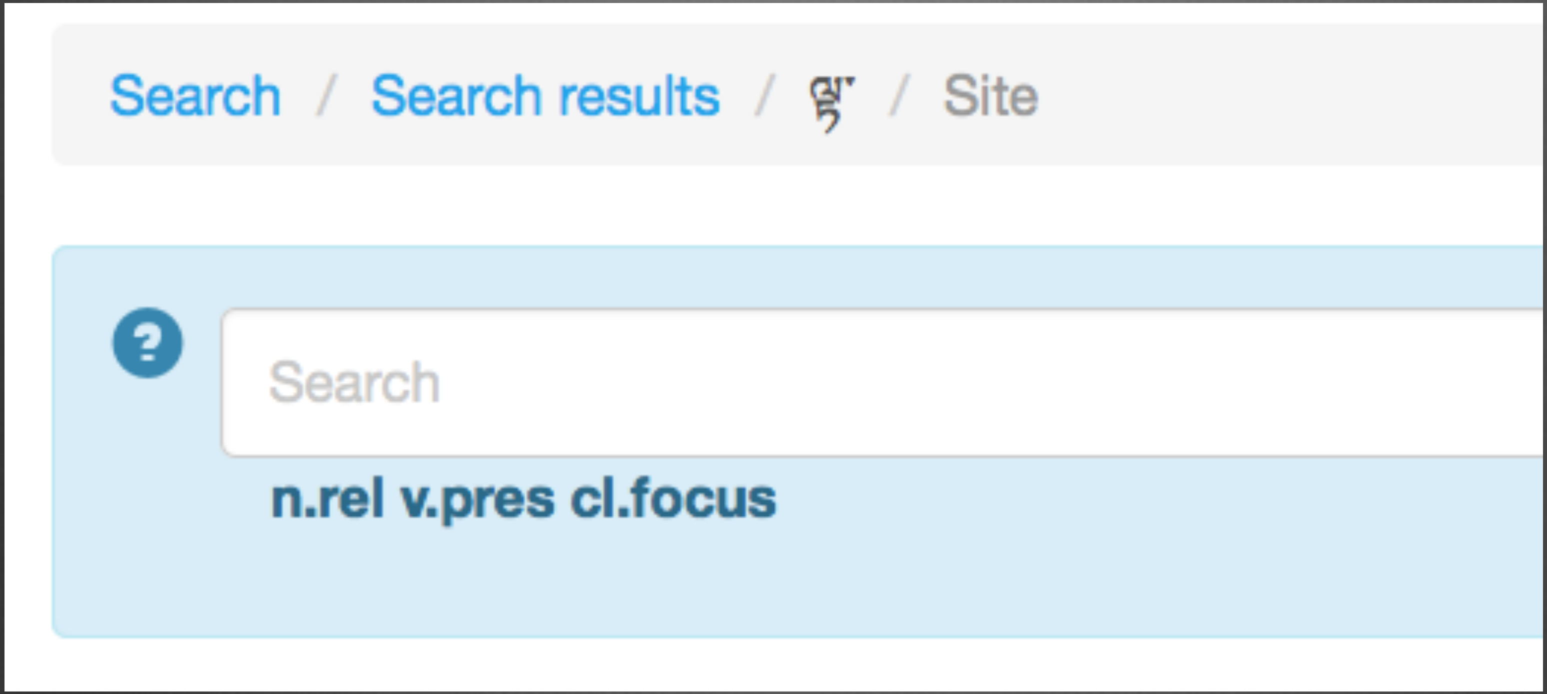
Word search

The website's search functionality is currently limited to exact match searching for Tibetan words. If you enter a Tibetan word, then the system will find all occurrences of the word, allowing you to further narrow your search by part-of-speech if the word form is ambiguous. For example, try typing ཐོག་ into the search box.

Shingle search

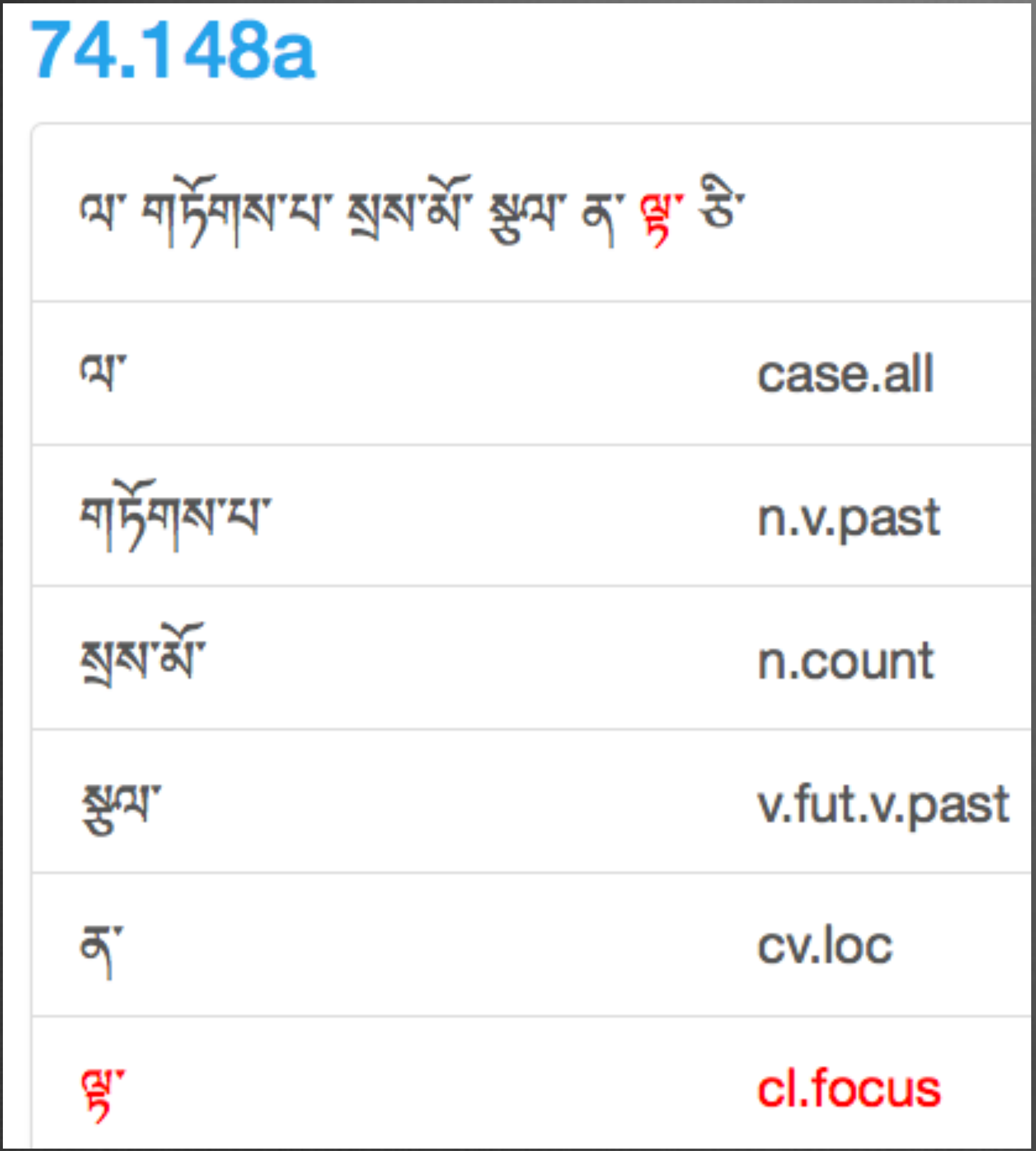
A second kind of searching helps to find interesting patterns in pos taggings. In the "shingle search" interface, whole corpora are tagged from scratch using our current best segmenter followed by the rule tagger. These search pages use a Flash plugin, ZeroClipboard, to copy the shingle tables to the clipboard, and to export them to CSV, Excel, or PDF formats. These functions won't work on mobile platforms and browsers lacking Flash.

Search



n.rel v.pres cl.focus

Search



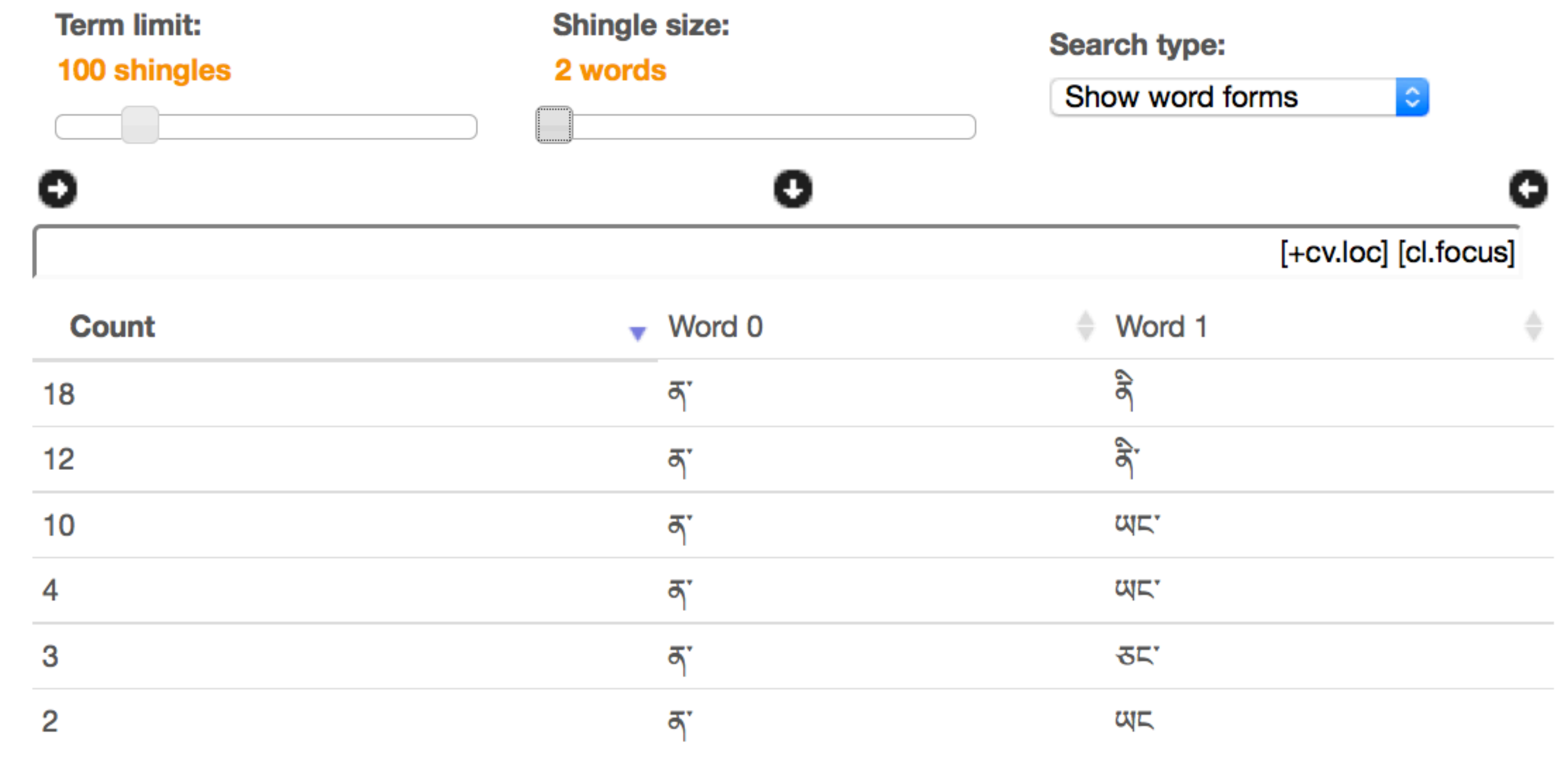
74.148a

ལ་ གཏོགས་པ་ སྲས་མོ་ སྐལ་ ན་ ལྟ་ ཅི་

ལ་	case.all
གཏོགས་པ་	n.v.past
སྲས་མོ་	n.count
སྐལ་	v.fut.v.past
ན་	cv.loc
ལྟ་	cl.focus

Shingles

Looking for [cl.focus] after [cv.loc]



Term limit:

100 shingles

Shingle size:

2 words

Search type:

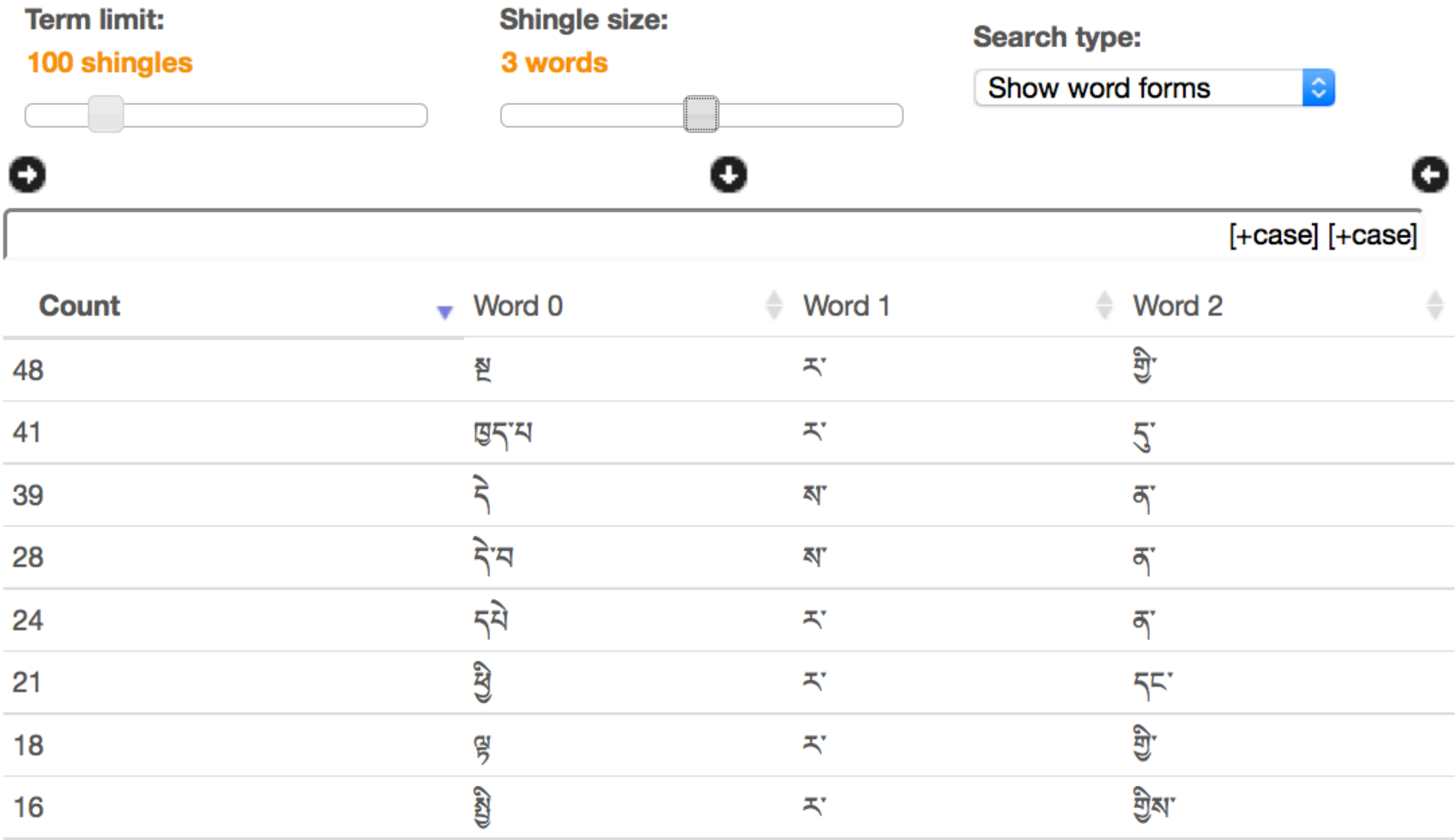
Show word forms

[+cv.loc] [cl.focus]

Count	Word 0	Word 1
18	ན་	ནི
12	ན་	ནི་
10	ན་	ཡང་
4	ན་	ཡང་
3	ན་	ཅང་
2	ན་	ཡང

Shingles

Looking for double case marking.



Term limit:
100 shingles

Shingle size:
3 words

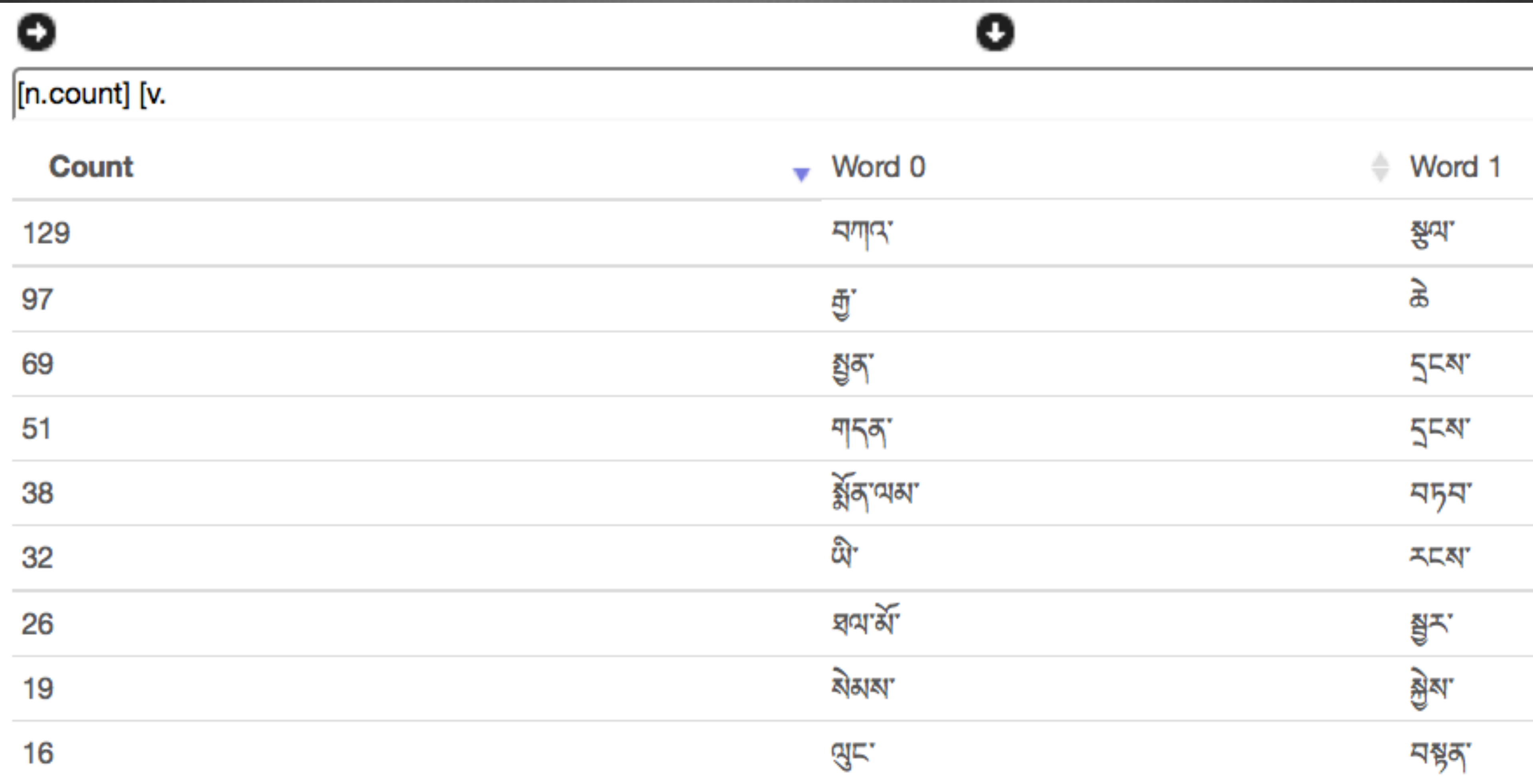
Search type:
Show word forms

[+case] [+case]

Count	Word 0	Word 1	Word 2
48	སྔ	ར་	གྱི་
41	ཁུད་པ	ར་	དུ་
39	དེ	ས་	ན་
28	དེ་བ	ས་	ན་
24	དཔེ	ར་	ན་
21	ཕྱི	ར་	དང་
18	ལྷ	ར་	གྱི་
16	སྤྱི	ར་	གྱིས་

Shingles

Common collocations.



[n.count] [v.

Count	Word 0	Word 1
129	བཀའ་	སྩལ་
97	རྒྱ་	ཆེ
69	སྤྱན་	དྲངས་
51	གདན་	དྲངས་
38	སྨོན་ལམ་	བཏབ་
32	ཡི་	རངས་
26	ཐལ་མོ་	སྦྱར་
19	སེམས་	སྐྱེས་
16	ལུང་	བསྟན་

Discovering new things about Tibetan grammar

Conclusions on infinitive constructions

1. Past tense verbs do not occur as the subordinate verbs of indirect infinitives.
2. The matrix verbs *gsol, med, grags, yod, ruṅ* select the future tense.
3. 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.

Garrett, Edward and Hill, Nathan W. and Zadoks, Abel (2013) 'Disambiguating Tibetan verb stems with matrix verbs in the indirect infinitive construction.' *Bulletin of Tibetology*, 49 (2). pp. 35-44.

How well does it work?
Accuracy and Ambiguity

Classical (159,144 words)	Accuracy	Ambiguity
LexTagger	1.00000	2.50755
RuleTagger	0.99906	1.37665
Difference		**1.13090**

(on 14 Nov 2014)

Classical (206,007 words)	Accuracy	Ambiguity
LexTagger	0.99999	2.63390
RuleTagger	0.99892	1.40948
Difference		**1.22442**

(on 05 March 2015)

Classical (226,021 words)	Accuracy	Ambiguity
LexTagger	1.00000	2.64819
RuleTagger	0.99901	1.40909
Difference		**1.23910**

(on 16 April 2015)

Thank you