Evidentiality Workshop, Paris

The Emergence of Egophoricity:

A Diachronic Investigation into the Marking of the Conscious Self

Christian Faggionato, Nathan Hill & Marieke Meelen
SOAS & Cambridge

25 February 2022

AHRC-funded project 2022-2026

The following research project is founded by AHRC, the Art and Humanities Research Council (UK) and will last till February 2026. The project members are:

Dr Marieke Meelen - University of Cambridge

Prof Nathan Hill - Trinity College Dublin

Dr Christian Faggionato - University of Cambridge

Dr Alexander James O'Neill - SOAS - London

Introduction

How do languages develop egophoric marking?

- Newar and Tibetan offer an excellent starting point to answer this question.

Introduction

How do languages develop egophoric marking?

- Newar and Tibetan offer an excellent starting point to answer this question.

- Tibetan and Newar have long literary traditions (Tibetan since 650 CE and Newar since 1112 CE). Unlike their present-day descendants, neither Classical Tibetan or Classical Newar exhibit egophoricity (Tournadre & Jiatso 2001)

Introduction

How do languages develop egophoric marking?

- Newar and Tibetan offer an excellent starting point to answer this question.
- Tibetan and Newar have long literary traditions (Tibetan since 650 CE and Newar since 1112 CE). Unlike their present-day descendants, neither Classical Tibetan or Classical Newar exhibit egophoricity (Tournadre & Jiatso 2001)
- This project will provide the first comprehensive diachronic quantitative & qualitative investigation of egophoric marking, by tracing the development of egophoricity in Tibetan and Newar through time and space.

Introduction

Newar and Tibetan mark the speaker ’ s personal involvement in an event:

In Lhasa Tibetan all sentences ending with ***yin*** involve the speaker somehow (examples from Tournadre & Dorje 2003):

(1) a. nga em-chi **yin**
I doctor YIN
‘**I**’m a doctor’

b. ’di nga’i bu-mo **yin**
this my daughter YIN
‘This is **my** daughter’

c. ’di khyed-rang-gi gsol-ja **yin**
this your tea YIN
‘This is your tea [that **I** have made for you]’

Introduction

Kathmandu Newar uses vowel lengthening to indicate a speaker's involvement but only if self-conscious:

(2) a. jĩ: a:pwa twan-**ā**
I too.much drank-**ā**
'I drank too much'

b. chã/wa a:pwa twan-**a**
you too.much drank-**a**
'you/(s)he drank too much'

(3) a. jĩ: Mānaj nāpalān-**ā**
I Manaj met-**ā**
'I met Manoj as planned'

b. jĩ: Mānaj nāpalān-**a**
I Manaj met-**a**
'I met Manoj by coincidence

Research Questions

How did they end up with such different egophoric systems? The main research questions are:

- Q1 - What type of synchronic variation can we find in egophoric marking in classical and present day Newar and Tibetan varieties?

Research Questions

How did they end up with such different egophoric systems? The main research questions are:

- Q1 - What type of synchronic variation can we find in egophoric marking in classical and present day Newar and Tibetan varieties?
- Q2 - What is the origin and diachronic development of egophoric markers in these varieties respect to their morphosyntactic, semantic and pragmatic dimensions?

Research Questions

How did they end up with such different egophoric systems? The main research questions are:

- Q1 - What type of synchronic variation can we find in egophoric marking in classical and present day Newar and Tibetan varieties?
- Q2 - What is the origin and diachronic development of egophoric markers in these varieties respect to their morphosyntactic, semantic and pragmatic dimensions?
- Q3 - How do these dimensions interact with each other?

Research Questions

How did they end up with such different egophoric systems? The main research questions are:

- Q1 - What type of synchronic variation can we find in egophoric marking in classical and present day Newar and Tibetan varieties?
- Q2 - What is the origin and diachronic development of egophoric markers in these varieties respect to their morphosyntactic, semantic and pragmatic dimensions?
- Q3 - How do these dimensions interact with each other?
- Q4 - What triggers of language change can we identify in the emergence of egophoricity?

Theoretical Perspectives & Methods

Our project combines a variety of empirical and theoretical perspectives and methods:

- Original fieldwork in Nepal (Newar & SMT)

Theoretical Perspectives & Methods

Our project combines a variety of empirical and theoretical perspectives and methods:

- Original fieldwork in Nepal (Newar & SMT)
- NLP methods for corpus building (including annotation)

Theoretical Perspectives & Methods

Our project combines a variety of empirical and theoretical perspectives and methods:

- Original fieldwork in Nepal (Newar & SMT)
- NLP methods for corpus building (including annotation)
- Synchronic theoretical syntax, semantics & pragmatics

Theoretical Perspectives & Methods

Our project combines a variety of empirical and theoretical perspectives and methods:

- Original fieldwork in Nepal (Newar & SMT)
- NLP methods for corpus building (including annotation)
- Synchronic theoretical syntax, semantics & pragmatics
- Language-internal & -external triggers for language change

Previous Diachronic Studies

- Some diachronic studies for Newar and Tibetan - semantic developments and grammaticalisation of single lexical items (Oisel 2013, Hargreaves 2018)

Previous Diachronic Studies

- Some diachronic studies for Newar and Tibetan - semantic developments and grammaticalisation of single lexical items (Oisel 2013, Hargreaves 2018)
- Some studies on historical comparative reconstructions of egophoric functions and systems as a whole in Newar and Tibetan (DeLancey 1992, Widmer & Zemp 2017, Zemp 2020)

Previous Diachronic Studies

- Some diachronic studies for Newar and Tibetan - semantic developments and grammaticalisation of single lexical items (Oisel 2013, Hargreaves 2018)
- Some studies on historical comparative reconstructions of egophoric functions and systems as a whole in Newar and Tibetan (DeLancey 1992, Widmer & Zemp 2017, Zemp 2020)
- Some general grammaticalisation pathways have been proposed for the emergence of egophoric marking or speaker-involvement in functional and formal historical linguistic literature (Traugott 1995, Nuckolls & Michael 2012, Roberts & Roussou 2003)

Annotated Historical Corpora

None of these strands of linguistic research have tested and substantiated their claims by empirical data from large-scale annotated historical corpora. Corpora will allow us to:

- Build on and complement descriptive/typological studies discussing synchronic and diachronic variation (Research Q1 & Q2) with new and comprehensive data

Annotated Historical Corpora

None of these strands of linguistic research have tested and substantiated their claims by empirical data from large-scale annotated historical corpora. Corpora will allow us to:

- Build on and complement descriptive/typological studies discussing synchronic and diachronic variation (Research Q1 & Q2) with new and comprehensive data
- Test theoretical hypotheses about interaction between morphosyntactic, semantic and pragmatic dimensions (Research Q3)

Annotated Historical Corpora

None of these strands of linguistic research have tested and substantiated their claims by empirical data from large-scale annotated historical corpora. Corpora will allow us to:

- Build on and complement descriptive/typological studies discussing synchronic and diachronic variation (Research Q1 & Q2) with new and comprehensive data
- Test theoretical hypotheses about interaction between morphosyntactic, semantic and pragmatic dimensions (Research Q3)
- Test ‘mechanisms’ and triggers of change (Research Q4)

Annotated Historical Corpora

Annotated Historical Corpora can facilitate a large amount of **morphosyntactic** and **information-structural** (i.e. semantic and pragmatic) research:

- Morphosyntactic information is indispensable to provide a comprehensive overview of the distribution and functions of suffixes, auxiliaries and independent verb forms and their arguments

Annotated Historical Corpora

Annotated Historical Corpora can facilitate a large amount of **morphosyntactic** and **information-structural** (i.e. semantic and pragmatic) research:

- Morphosyntactic information is indispensable to provide a comprehensive overview of the distribution and functions of suffixes, auxiliaries and independent verb forms and their arguments
- Semantic annotation of predicate types (volitional/control, animacy) provides insights into the distribution of egophoric markers

Annotated Historical Corpora

Annotated Historical Corpora can facilitate a large amount of **morphosyntactic** and **information-structural** (i.e. semantic and pragmatic) research:

- Morphosyntactic information is indispensable to provide a comprehensive overview of the distribution and functions of suffixes, auxiliaries and independent verb forms and their arguments
- Semantic annotation of predicate types (volitional/control, animacy) provides insights into the distribution of egophoric markers
- Information-structural (pragmatic) annotation can indicate where the information comes from, whether it is new or old, and how a sentence is anchored in context, e.g. switch reference & topic chains

Research methods

The project will use a combination of fieldwork, descriptive, grammar- and discourse-oriented methods. Data Collection and Processing necessarily precede our analysis, which falls into the following strands:

- **Strand 1**: Synchronic variation in form & function of egophoric systems at different stages of Tibetan & Newar history

Research methods

The project will use a combination of fieldwork, descriptive, grammar- and discourse-oriented methods. Data Collection and Processing necessarily precede our analysis, which falls into the following strands:

- **Strand 1**: Synchronic variation in form & function of egophoric systems at different stages of Tibetan & Newar history
- **Strand 2**: Diachronic change of Tibetan Newar egophoric markers in terms of morphosyntax, semantics and pragmatics

Research methods

The project will use a combination of fieldwork, descriptive, grammar- and discourse-oriented methods. Data Collection and Processing necessarily precede our analysis, which falls into the following strands:

- **Strand 1**: Synchronic variation in form & function of egophoric systems at different stages of Tibetan & Newar history
- **Strand 2**: Diachronic change of Tibetan Newar egophoric markers in terms of morphosyntax, semantics and pragmatics
- **Strand 3**: The interaction of morphosyntax, semantics and pragmatics

Research methods

The project will use a combination of fieldwork, descriptive, grammar- and discourse-oriented methods. Data Collection and Processing necessarily precede our analysis, which falls into the following strands:

- **Strand 1**: Synchronic variation in form & function of egophoric systems at different stages of Tibetan & Newar history
- **Strand 2**: Diachronic change of Tibetan Newar egophoric markers in terms of morphosyntax, semantics and pragmatics
- **Strand 3**: The interaction of morphosyntax, semantics and pragmatics
- **Strand 4**: 'Mechanisms' and triggers of change in egophoric markers over time and in crosslinguistic perspective: within the Tibeto-Burman language family and beyond

Data Collection

- Data collection and selection is driven by considerations of genre, representativeness and, availability

Data Collection

- Data collection and selection is driven by considerations of genre, representativeness and, availability

- Protocols for automatic segmentation, POS tagging and parsing have already been developed for Tibetan (Meelen & Hill 2017), but information-structural/pragmatic annotation and correction of morphosyntactic annotation require labour intensive extension and correction –> **so we develop new semi-supervised methods**

Data Collection

- Data collection and selection is driven by considerations of genre, representativeness and, availability
- Protocols for automatic segmentation, POS tagging and parsing have already been developed for Tibetan (Meelen & Hill 2017), but information-structural/pragmatic annotation and correction of morphosyntactic annotation require labour intensive extension and correction –> **so we develop new semi-supervised methods**
- First-hand data collection through fieldwork will be limited to oral narratives in the Lalitput Newar variety with our local consultants

Types of Data and Methods

The project produces two digital resources:

1. An **automatically** annotated corpus of 200m Tibetan and 180k Newar tokens; for Tibetan, existing morphosyntactically annotated corpora largely meet our needs. For Newar more preprocessing is needed.

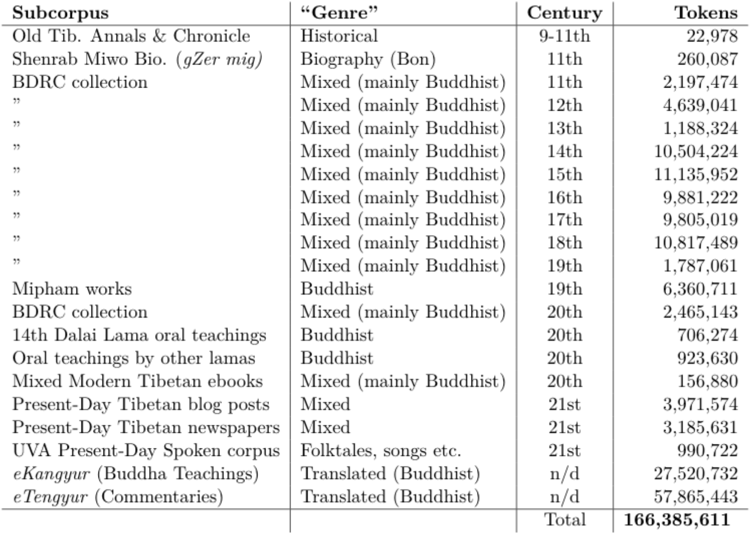
Types of Data and Methods

The project produces two digital resources:

1. An **automatically** annotated corpus of 200m Tibetan and 180k Newar tokens; for Tibetan, existing morphosyntactically annotated corpora largely meet our needs. For Newar more preprocessing is needed.

2. A **manually**, **deeply annotated** corpus of 240k Tibetan and 180k Newar tokens: includes annotation of specific information-structural (including speech act) labels that cannot be generated automatically (e.g. givenness, speech context, etc)

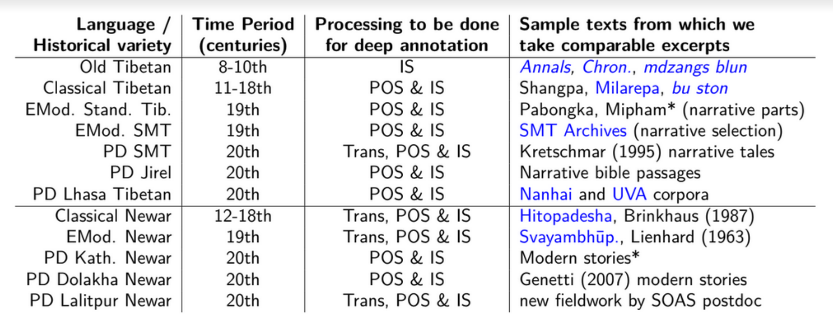
The PArsed Corpus of Tibetan (Output a)



Subcorpus	**“Genre”**	**Century**	**Tokens**
Old Tib. Annals & Chronicle	Historical	9-11th	22,978
Shenrab Miwo Bio. (*gZer mig*)	Biography (Bon)	11th	260,087
BDRC collection	Mixed (mainly Buddhist)	11th	2,197,474
”	Mixed (mainly Buddhist)	12th	4,639,041
”	Mixed (mainly Buddhist)	13th	1,188,324
”	Mixed (mainly Buddhist)	14th	10,504,224
”	Mixed (mainly Buddhist)	15th	11,135,952
”	Mixed (mainly Buddhist)	16th	9,881,222
”	Mixed (mainly Buddhist)	17th	9,805,019
”	Mixed (mainly Buddhist)	18th	10,817,489
”	Mixed (mainly Buddhist)	19th	1,787,061
Mipham works	Buddhist	19th	6,360,711
BDRC collection	Mixed (mainly Buddhist)	20th	2,465,143
14th Dalai Lama oral teachings	Buddhist	20th	706,274
Oral teachings by other lamas	Buddhist	20th	923,630
Mixed Modern Tibetan ebooks	Mixed (mainly Buddhist)	20th	156,880
Present-Day Tibetan blog posts	Mixed	21st	3,971,574
Present-Day Tibetan newspapers	Mixed	21st	3,185,631
UVA Present-Day Spoken corpus	Folktales, songs etc.	21st	990,722
eKangyur (Buddha Teachings)	Translated (Buddhist)	n/d	27,520,732
eTengyur (Commentaries)	Translated (Buddhist)	n/d	57,865,443
		Total	**166,385,611**

Additional Deeply Annotated Data (Output b)

We aim to create balanced deeply annotated corpora:

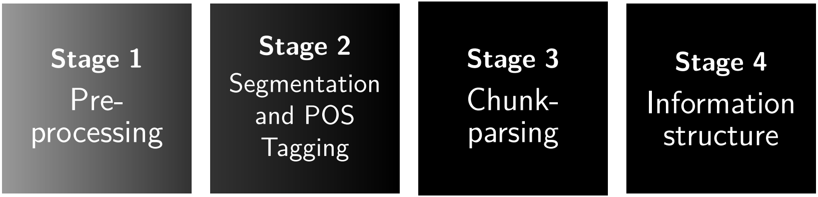


Language / Historical variety	Time Period (centuries)	Processing to be done for deep annotation	Sample texts from which we take comparable excerpts
Old Tibetan	8-10th	IS	*Annals*, *Chron.*, *mdzangs blun*
Classical Tibetan	11-18th	POS & IS	Shangpa, Milarepa, *bu ston*
EMod. Stand. Tib.	19th	POS & IS	Pabongka, Mipham* (narrative parts)
EMod. SMT	19th	POS & IS	SMT Archives (narrative selection)
PD SMT	20th	Trans, POS & IS	Kretschmar (1995) narrative tales
PD Jirel	20th	POS & IS	Narrative bible passages
PD Lhasa Tibetan	20th	POS & IS	Nanhai and UVA corpora
Classical Newar	12-18th	Trans, POS & IS	Hitopadesha, Brinkhaus (1987)
EMod. Newar	19th	Trans, POS & IS	Svayambhūp., Lienhard (1963)
PD Kath. Newar	20th	POS & IS	Modern stories*
PD Dolakha Newar	20th	POS & IS	Genetti (2007) modern stories
PD Lalitpur Newar	20th	Trans, POS & IS	new fieldwork by SOAS postdoc

EMod. = Early Modern; PD = Present-Day; SMT = South Mustang Tibetan; IS = Information-structural/pragmatic annotation; POS = part-of-speech/morphosyntactic annotation; Trans. = transcription; * = from personal collections

First Results: Annotation Workflow

First, we need to develop, test and implement an Annotation Workflow:



Stage 1
Pre-processing

Stage 2
Segmentation and POS Tagging

Stage 3
Chunk-parsing

Stage 4
Information structure

Pre-processing

Pre-processing involves the following tasks:

1. Download Unicode text from ODTO and create .txt file

Pre-processing

Pre-processing involves the following tasks:

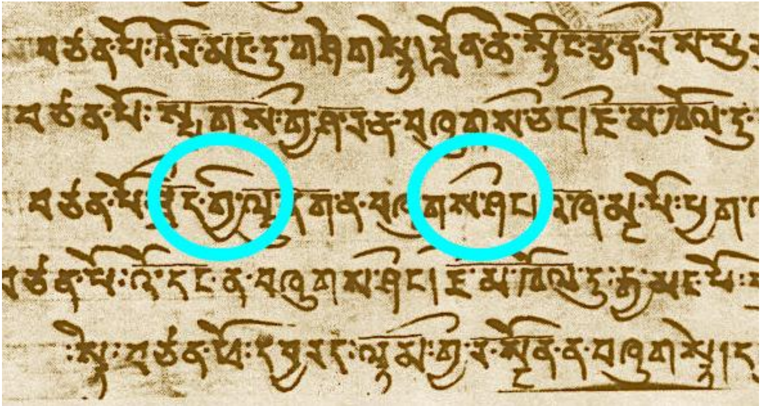
1. Download Unicode text from ODTO and create .txt file
2. Cleaning editorial conventions with Regular Expression

Pre-processing

Pre-processing involves the following tasks:

1. Download Unicode text from ODTO and create .txt file
2. Cleaning editorial conventions with Regular Expression
3. Normalise Old Tibetan > Classical Tibetan with CG3

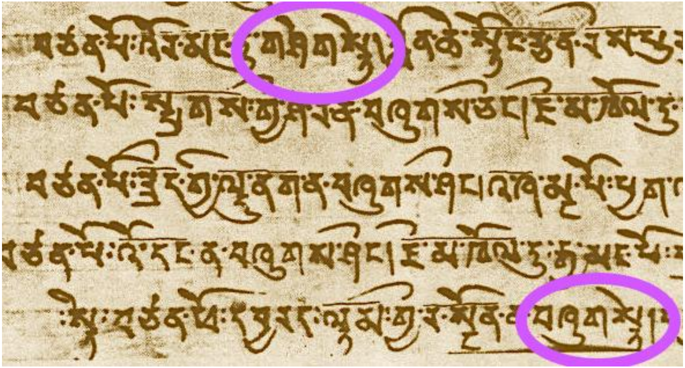
Pre-processing - Normalisation of vowel marks



ཤྀང་ > ཤིང་

sh-ing > *shing*

Pre-processing - Normalisation of merged syllables



གཤེགསྟེ་ > གཤེགས་ཏེ་
བཞུགསྟེ་ > བཞུགས་ཏེ་

gshegste > *gshegs te*
bzhugste > *bzhugs te*

Normalisation with Cg3

Splitting merged syllables like

བཞུགསོ་ > བཞུགས་སོ་ *bzhugso* > *bzhugs so*

...with a Cg3 grammar:

```
SPLITCOHORT ( "<$1>"v "$1$3 "v "<$3$4>"v "$3$4"v )
("<(.2,6)(([^\u0FB2\u0FB1])

([\u0F7C\u0F7A\u0F74\u0F72\u0F80]
 ?))>"r)(NOT 0 (split) or (genitive) or (diphthongs));
```

Segmentation and POS Tagging

Segmentation and POS Tagging inolve the following tasks:

1. Automatic Word Segmentation (+ Rule-based correction)

Segmentation and POS Tagging

Segmentation and POS Tagging inolve the following tasks:

1. Automatic Word Segmentation (+ Rule-based correction)
2. Automatic POS Tagging

Segmentation and POS Tagging

Segmentation and POS Tagging inolve the following tasks:

1. Automatic Word Segmentation (+ Rule-based correction)
2. Automatic POS Tagging
3. Automatic Sentence Segmentation

Segmentation and POS Tagging

	Global Accuracy
Classical Tibetan (318*k*; 15 tags)	96.3%
Old Tibetan (3.5*k*; 15 tags)	92.8%
Old & Classical (321.5*k*; 15 tags)	96.1%
Wylie transliteration (318*k*; 15 tags)	**96.5%**
Unicode Tibetan (318*k*; 79 tags)	95.0%
Wylie transliteration (318*k*; 79 tags)	94.7%
Neural-Network tagger (318*k*; 79 tags)	**95.8%**

Segmentation and POS Tagging

	Global Accuracy
Classical Tibetan (318*k*; 15 tags)	96.3%
Old Tibetan (3.5*k*; 15 tags)	92.8%
Old & Classical (321.5*k*; 15 tags)	96.1%
Wylie transliteration (318*k*; 15 tags)	**96.5%**
Unicode Tibetan (318*k*; 79 tags)	95.0%
Wylie transliteration (318*k*; 79 tags)	94.7%
Neural-Network tagger (318*k*; 79 tags)	**95.8%**

- Small tag set, Wylie transliteration & Memory-Based tagger best so far...

Segmentation and POS Tagging

After the automated segmentation and POS tagging some manual corrections are required through Pyrrha - a webapp for fast and secure morphological post-correction or annotation:

1. Create Pyrrha input format

Segmentation and POS Tagging

After the automated segmentation and POS tagging some manual corrections are required through Pyrrha - a webapp for fast and secure morphological post-correction or annotation:

1. Create Pyrrha input format
2. Manual correction in Pyrrha of Word segmentation, Sentence segmentation & POS tagging

Segmentation and POS Tagging

After the automated segmentation and POS tagging some manual corrections are required through Pyrrha - a webapp for fast and secure morphological post-correction or annotation:

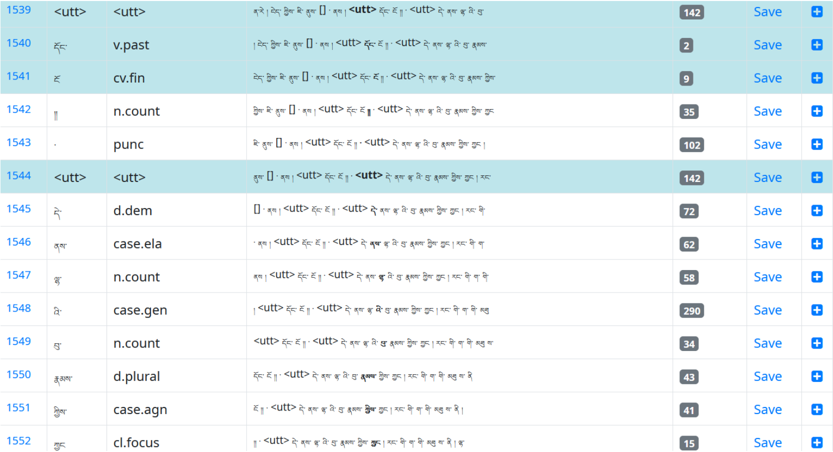
1. Create Pyrrha input format
2. Manual correction in Pyrrha of Word segmentation, Sentence segmentation & POS tagging
3. Create utterance boundaries (based on verbs, particles, etc.)

Segmentation and POS Tagging

After the automated segmentation and POS tagging some manual corrections are required through Pyrrha - a webapp for fast and secure morphological post-correction or annotation:

1. Create Pyrrha input format
2. Manual correction in Pyrrha of Word segmentation, Sentence segmentation & POS tagging
3. Create utterance boundaries (based on verbs, particles, etc.)
4. Add SentenceIDs + automatic POS replacements

Segmentation and POS Tagging: Pyrrha correction



1539	<utt>	<utt>	ན་རེ ། དེད་ ཀྱིས་ ཇི་ ནུས་ [] · ནས ། **<utt>** སོང་ སོ ༎ · <utt> དེ་ ནས་ ལྷ་ འི་ བུ་	142	Save	+
1540	སོང་	v.past	། དེད་ ཀྱིས་ ཇི་ ནུས་ [] · ནས ། <utt> **སོང་** སོ ༎ · <utt> དེ་ ནས་ ལྷ་ འི་ བུ་ རྣམས་	2	Save	+
1541	སོ	cv.fin	དེད་ ཀྱིས་ ཇི་ ནུས་ [] · ནས ། <utt> སོང་ **སོ** ༎ · <utt> དེ་ ནས་ ལྷ་ འི་ བུ་ རྣམས་ ཀྱིས་	9	Save	+
1542	༎	n.count	ཀྱིས་ ཇི་ ནུས་ [] · ནས ། <utt> སོང་ སོ **༎** · <utt> དེ་ ནས་ ལྷ་ འི་ བུ་ རྣམས་ ཀྱིས་ ཀྱང	35	Save	+
1543	·	punc	ཇི་ ནུས་ [] · ནས ། <utt> སོང་ སོ ༎ · <utt> དེ་ ནས་ ལྷ་ འི་ བུ་ རྣམས་ ཀྱིས་ ཀྱང །	102	Save	+
1544	<utt>	<utt>	ནུས་ [] · ནས ། <utt> སོང་ སོ ༎ · **<utt>** དེ་ ནས་ ལྷ་ འི་ བུ་ རྣམས་ ཀྱིས་ ཀྱང ། རང་	142	Save	+
1545	དེ་	d.dem	[] · ནས ། <utt> སོང་ སོ ༎ · <utt> **དེ་** ནས་ ལྷ་ འི་ བུ་ རྣམས་ ཀྱིས་ ཀྱང ། རང་ གི་	72	Save	+
1546	ནས་	case.ela	· ནས ། <utt> སོང་ སོ ༎ · <utt> དེ་ **ནས་** ལྷ་ འི་ བུ་ རྣམས་ ཀྱིས་ ཀྱང ། རང་ གི་ ག་	62	Save	+
1547	ལྷ་	n.count	ནས ། <utt> སོང་ སོ ༎ · <utt> དེ་ ནས་ **ལྷ་** འི་ བུ་ རྣམས་ ཀྱིས་ ཀྱང ། རང་ གི་ ག་ གི་	58	Save	+
1548	འི་	case.gen	། <utt> སོང་ སོ ༎ · <utt> དེ་ ནས་ ལྷ་ **འི་** བུ་ རྣམས་ ཀྱིས་ ཀྱང ། རང་ གི་ ག་ གི་ མཐུ	290	Save	+
1549	བུ་	n.count	<utt> སོང་ སོ ༎ · <utt> དེ་ ནས་ ལྷ་ འི་ **བུ་** རྣམས་ ཀྱིས་ ཀྱང ། རང་ གི་ ག་ གི་ མཐུ ས་	34	Save	+
1550	རྣམས་	d.plural	སོང་ སོ ༎ · <utt> དེ་ ནས་ ལྷ་ འི་ བུ་ **རྣམས་** ཀྱིས་ ཀྱང ། རང་ གི་ ག་ གི་ མཐུ ས་ ནི	43	Save	+
1551	ཀྱིས་	case.agn	སོ ༎ · <utt> དེ་ ནས་ ལྷ་ འི་ བུ་ རྣམས་ **ཀྱིས་** ཀྱང ། རང་ གི་ ག་ གི་ མཐུ ས་ ནི །	41	Save	+
1552	ཀྱང	cl.focus	༎ · <utt> དེ་ ནས་ ལྷ་ འི་ བུ་ རྣམས་ ཀྱིས་ **ཀྱང** ། རང་ གི་ ག་ གི་ མཐུ ས་ ནི ། ལྷ་	15	Save	+

Parsing

Parsing involves the following tasks:

- Develop rule-based grammar for Tibetan and Newar

Parsing

Parsing involves the following tasks:

- Develop rule-based grammar for Tibetan and Newar
- Automatic (hierarchical) chunk-parsing using NLTK RegEx parser

Parsing

Parsing involves the following tasks:

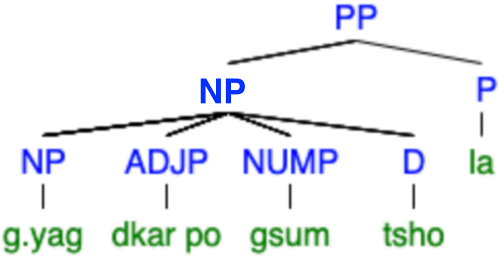
- Develop rule-based grammar for Tibetan and Newar
- Automatic (hierarchical) chunk-parsing using NLTK RegEx parser
- Manual parse correction with Cesax

Parsing - Postpositional Phrases

གཡག་དཀར་པོ་གསུམ་ཚོ་ལ

g.yag dkar po gsum tsho la
“to the three white yaks”

(PP (NP (NP *g.yag* (ADJP *dkar po*) (NUMP *gsum*) D *tsho*)) ADP *la*)



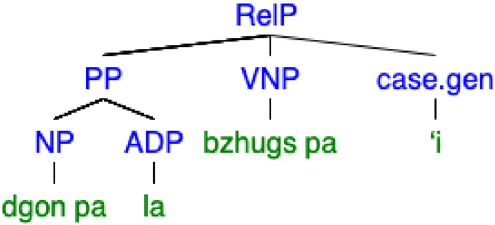
Parsing - Relative Clauses

དགོན་པ་ན་བཞུགས་པའི་བླ་མ

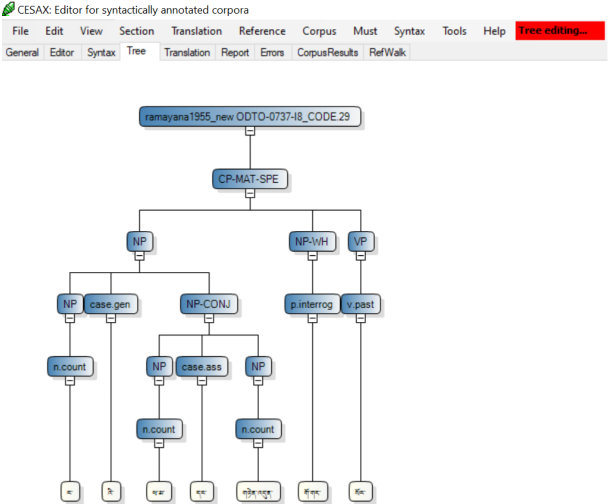
dgon pa na bzhugs pa'i *bla ma*

"the lama **who dwells at the monastery**"

(RelP (PP (NP *dgon pa*) (ADP *la*)) (VNP *bzhugs pa*) (case.gen ' *i*)



Manual Parse Correction with Cesax



Information Structure Annotation

Information Structure involves the following tasks:

1. Sentence types (declarative/questions/direct speech, etc.)

Information Structure Annotation

Information Structure involves the following tasks:

1. Sentence types (declarative/questions/direct speech, etc.)
2. Topics (Aboutness/Familiar/Contrastive)

Information Structure Annotation

Information Structure involves the following tasks:

1. Sentence types (declarative/questions/direct speech, etc.)
2. Topics (Aboutness/Familiar/Contrastive)
3. Focus Domain (Constituent/Predicate/Presentational)

Information Structure Annotation

Information Structure involves the following tasks:

1. Sentence types (declarative/questions/direct speech, etc.)
2. Topics (Aboutness/Familiar/Contrastive)
3. Focus Domain (Constituent/Predicate/Presentational)
4. Animacy: create word embedding-based classifier for human/animate/inanimate

Information Structure Annotation

Information Structure involves the following tasks:

1. Sentence types (declarative/questions/direct speech, etc.)
2. Topics (Aboutness/Familiar/Contrastive)
3. Focus Domain (Constituent/Predicate/Presentational)
4. Animacy: create word embedding-based classifier for human/animate/inanimate
5. Verbal features (volitional/non-volitional etc.)

Information Structure: Topics

Topics can detected semi-automatically:

1. Find *ni* marker (tagged with `cl.top`)

Information Structure: Topics

Topics can detected semi-automatically:

1. Find *ni* marker (tagged with `cl.top`)
2. Combine with previous consituent into TopicPhrase

Information Structure: Topics

Topics can detected semi-automatically:

1. Find *ni* marker (tagged with `cl.top`)
2. Combine with previous consituent into TopicPhrase
3. Add feature specifying type of topic (semi-automatically)

Information Structure: Focus

Focus domains can be detected semi-automatically:

1. Find *yang/kyang/...* markers (tagged with `cl.focus`)

Information Structure: Focus

Focus domains can be detected semi-automatically:

1. Find *yang/kyang/...* markers (tagged with `cl.focus`)
2. If preceding phrase is NP/PP, create NP-FOC/PP-FOC (`Constituent Focus`)

Information Structure: Focus

Focus domains can be detected semi-automatically:

1. Find *yang/kyang/...* markers (tagged with `cl.focus`)
2. If preceding phrase is NP/PP, create NP-FOC/PP-FOC (`Constituent Focus`)
3. Wh-questions are `Constituent Focus` too

Information Structure: Focus

Focus domains can be detected semi-automatically:

1. Find *yang/kyang/...* markers (tagged with `cl.focus`)
2. If preceding phrase is NP/PP, create NP-FOC/PP-FOC (`Constituent Focus`)
3. Wh-questions are `Constituent Focus` too
4. `Presentational focus` can be detected through certain verbs

Information Structure: Focus

Focus domains can be detected semi-automatically:

1. Find *yang/kyang/...* markers (tagged with `cl.focus`)
2. If preceding phrase is NP/PP, create NP-FOC/PP-FOC (`Constituent Focus`)
3. Wh-questions are `Constituent Focus` too
4. `Presentational focus` can be detected through certain verbs
5. Rest/default is `Predicate Focus`

Word embeddings

- Word embeddings enhance performance of NLP tools, so can we create them based on our segmented corpus?

Word embeddings

- Word embeddings enhance performance of NLP tools, so can we create them based on our segmented corpus?

- Yes, for Kanjur+Tengyur 'Can-Vec' (74m tokens) and for all of PACTib 'BDRC-Vec' (166m tokens): are they any good?

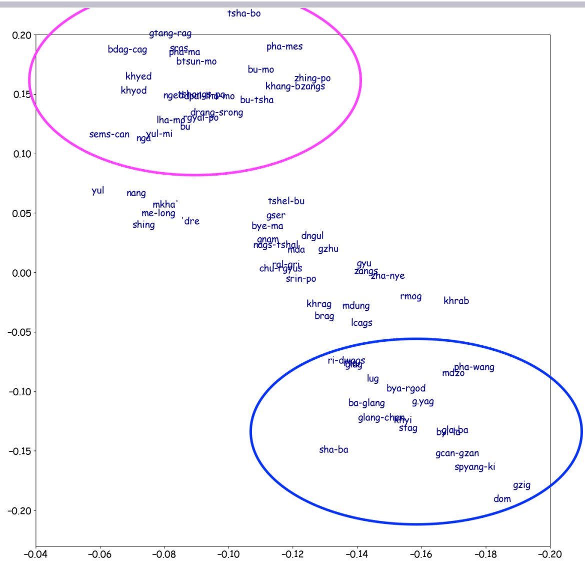
Word embeddings

- Word embeddings enhance performance of NLP tools, so can we create them based on our segmented corpus?

- Yes, for Kanjur+Tengyur ‘Can-Vec’ (74m tokens) and for all of PACTib ‘BDRC-Vec’ (166m tokens): are they any good?

- PACTib embeddings are boosting POS tagging from 93>97%

Word Embeddings for Animacy Detection



Quantitative Results & Hypothesis Testing

Once we have our corpora, we can answer a test a number of hypotheses to do with change over time:

- Which subjects go with which types of verbs?
- Which auxiliaries and verbal endings go with which type of sentence?
- How do *ni* and *kyis* interact?
- Does animacy play a role in sentences with *-nas* vs *-pa dang*?
- What role do copulas and resultatives play?
- Are there different trends for tense/aspect and reported speech?

Quantitative Results & Hypothesis Testing

These will lead to tests of broader hypotheses and questions:

- What information-structural features interact with which egophoric or evidential markers?
- Can we identify any triggers of change?
- What, if anything, can switch reference tell us?
- When, how and why do different varieties diverge?

Old Tibetan Ramayana: Sentence Types

We've tested our entire workflow from preprocessing to the annotation of information structure. Some very crude observations:

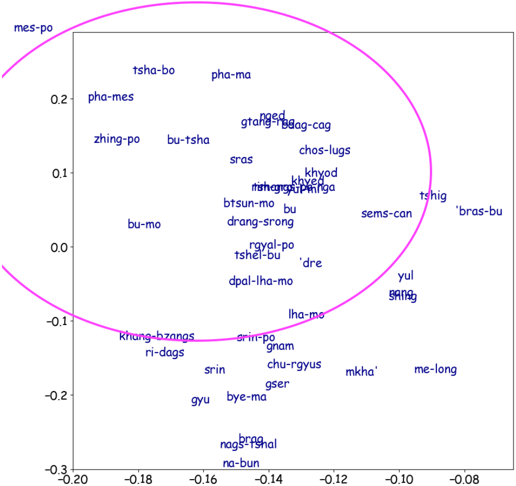
Type	Non-direct speech	Direct Speech
Matrix	29	29
Semi-final	5	10
cing	5	6
Quotative markers	12	n/a
Questions	n/a	3
Subordinates	31	n/a

Old Tibetan Ramayana: Agents/Subjects & Topics

Where do we find agents or subjects and topics?

Type	Main clause	Subordinate clause	Direct speech
NP-SBJ	4	3	5
NP-TOP	10	1	2

Old Tibetan Ramayana: Noun Animacy



Note that all human NPs are nicely clustered.

But there are not enough animal NPs in the Ramayana to form a meaningful group.

ཐུགས་རྗེ་ཆེ་

Thank you!

References

- Faggionato, C. & Meelen, M. 2019. Developing the Old Tibetan Treebank. In *Proceedings of the RANLP*.
- DeLancey, S. 1992. The hist. status of the conj/disj pattern. In *Acta Linguist. Hafniensa* 25:289–321.
- Hargreaves, D., 2018. Am I blue?. In Floyd et al (eds) *Egophoricity*, 118, 79-107.
- Meelen, M. & Hill, N. 2017. Segmenting and POS tagging Classical Tibetan. In *Himalayan Linguistics* 16 (2), 64-89.
- Nuckolls, J, and L. Michael. (eds) 2012. *Evidentiality in interaction*. Amsterdam: John Benjamins.
- Oisel, G., 2013. *Morphosyntaxe et sémantique des auxiliaires et des connecteurs du tibétain littéraire: étude diachronique et synchronique*. (Doctoral dissertation, Paris 3).
- Roberts, I. Roussou, A., 2003. *Syntactic change: A minimalist approach to grammaticalization*. CUP.
- Tournadre, N. & Jiatso, K., 2001. Final auxiliary verbs in literary Tibetan and in the dialects. In *Linguist. of the Tibeto-Burman Area*, 24(1), 49-111.
- Tournadre, N. & Dorje S., 2003. *Manual Of Standard Tibetan: Language And Civilization*. New York: Snow Lion Publications.
- Traugott, E.C., 1995. Subjectification in grammaticalization. In *Subjectivity and subjectivisation: Linguistic perspectives*, 1, 31-54.
- Widmer, M., 2017. The evolution of egophoricity and evidentiality in the Himalayas: The case of Bunan. In *J. of Hist. Linguist.*, 7(1), 245-274.
- Zemp, M., 2020. Evidentials and their pivot in Tibetic and neighboring Himalayan languages. In *Funct. of Lang.*, 27(1), 29-54.