# Meta-dating the PArsed Corpus of Tibetan (PACTib)

**Marieke Meelen**
University of Cambridge
Trinity Hall, Trinity Lane
Cambridge, UK, CB2 1TJ
mm986@cam.ac.uk

**Élie Roux**
Buddhist Digital Resource Center
1430 Massachusetts Ave., 5th floor
Cambridge, MA, USA 02138
roux.elie@gmail.com

## Abstract

This paper presents PACTib, the PArsed Corpus of Tibetan. This new resource is unique in bringing together a large number of Tibetan texts (>5000) from the 11th century until the present day. The texts in this diachronic corpus are provided with metadata containing information on dates and patron-/authorship and linguistic annotation in the form of tokenisation, sentence segmentation, part-of-speech tags and syntactic phrase structure. With over 166 million tokens across 11 centuries and a variety of genres, PACTib will open up a wide range of research opportunities for historical and comparative linguistics and scholars in Tibetan Studies, which we illustrate with two short case studies.

## 1 Introduction

In recent years a large number of Tibetan manuscripts and books have been digitised and electronic texts (manually transcribed or corrected after OCR/HTR) have been made available online by the Old Tibetan Documents Online project (OTDO), the Buddhist Digital Resource Center (BDRC) and Esukhia. In addition to these historical Tibetan texts, modern written Tibetan etexts can now be found on the websites of the Timeless Treasuries and Tibetan e-books initiatives, which include a mixture of genres and styles from around 1980 until today. Finally, a collection of songs, folktales and other oral narratives in Present-Day Spoken Tibetan was transcribed and deposited on Zenodo as the 'University of Virginia' (UVA) corpus. Despite the recent growth in digitised text materials, from an NLP point Tibetan is still an under-resourced and under-researched language. Most Tibetan NLP research to date has been carried out in China. However, the resulting publications[1] rarely make data or code available, effectively making it impossible to test, verify or use the results in any way. Instead, for the development of PACTib, we build on recent work on segmenting and POS tagging Tibetan by Garrett et al. (2014), Meelen and Hill (2017) and Faggionato and Meelen (2019) (see Section 3). In Section 2 we discuss the composition of the corpus and a proposal to allow for distinguishing easily between prose and verse. Section 3 focuses on the linguistic annotation. In Section 4 we add a brief note on how the texts in the corpus are linked to the relevant metadata. Finally, Section 5 presents two short case studies to illustrate the use of this unique historical treebank of Tibetan.

## 2 Composition of the annotated corpus

PACTib consists of a variety of digitised materials that have been made available online. For the historical materials (up to the 21st century), we initially only selected texts that were originally composed in Tibetan. We furthermore included texts containing teachings of the Buddha and commentaries on those (so-called *eKangyur* and *eTengyur* collections respectively) that were generally translated from Indic languages into Tibetan. The first witnesses of these translated texts sometimes date back to the 10th century. The digitised versions available today, however, are based on an 18th century edition, in which they have been substantially revised and edited.

[1] e.g. Liu et al. (2011)

Because of these issues with uncertain dates of origin (including revisions) and the fact that they are not originally composed in Tibetan, both the *eKangyur* and *eTengyur* collections are kept separate from the rest of the PACTib subcorpora in the results of the diachronic case studies (see Section 5). For comparative purposes, however, and because these texts are intensely studied by Buddhist scholars, we do include them in PACTib as it could be of interest to Tibetan Studies scholars studying these canonical texts and to linguists looking at potential issues of translated versus native Tibetan texts.

| **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** |

Table 1: Overview of PACTib Subcorpora

Table 1 gives an overview of all subcorpora that are currently included in the PACTib. The second column provisionally labelled "Genre" provides a rough indication of the type of texts contained in the subcorpora. The *Annals* and *Chronicle* are the earlier substantive amounts of Tibetan writing found in the Dunhuang caves in Gansu (Western China). These caves were sealed off in the 11th century and all manuscripts found in the caves are referred as 'Old Tibetan', the language spoken in the Yarlung Valley from where the Tibetan empire started its initial expansion. Most Old Tibetan texts are short inscriptions or more fragmentary parts of manuscripts and blockprints, but the *Annals* and *Chronicle* are longer and show more linguistic variety. Philologists generally consider the *Annals*, that record historical events in the 7-8th centuries, to be older than the more extensive *Chronicle*, although exact dates of origin are still a matter of ongoing debate (cf. Faggionato and Meelen (2019)). Tibetan texts written between the 11th and mid-20th centuries are generally referred to as 'Classical Tibetan', without further chronological subclassification. The two-volume biography of Shenrab Miwo (the founder of the Bon, i.e. a religion preceding Buddhism in Tibet) goes back to the 11th century, but is kept separate from the Old and Classical Tibetan texts since Bon texts generally contain non-Tibetan vocabulary as well (Snellgrove, 1967, 10). No systematic studies on differences in grammatical features have been done yet, although Snellgrove (1967, 8-9) makes some general remarks on the frequent mixing of genitive/agentive, locative/elative and allative/ablative case markers in Bon

texts in particular. The selection of electronic texts from the BDRC contain a wide variety of Buddhist writings in a range of topics from philosophy to religious teachings and commentaries, prayers in verse, ritual texts, songs and sometimes even novels dating from the 11th to the 20th centuries. It is important to note that in the texts from the BDRC collection not all centuries are equally well-represented: the amount of data from the 11th, 13th, 19th and 20th centuries in particular is rather low compared to other centuries. For this reason, we have decided to supplement the data for those centuries with texts from other sources as best as we could. For the 11th century data remains scarce in general and Shenrab Miwo's Bon Biography may not be the best point of comparison with the rest of the texts that are overwhelmingly Buddhist.[2] For the 13th century, there are no other sources we could use and therefore this century remains significantly underrepresented. When doing diachronic research, it is important to bear this in mind, in particular when aberrant patterns are found in the results from the 13th century.

The data from the 19th century could be supplemented by the works of the prolific Buddhist philosopher Mipham Jamyang Namgyal Gyamtso (1846–1912), who wrote over 32 volumes on a variety of topics such as poetics, sculpture, medicine, tantra and logic, digitised by Adarsha. Finally, from the 20th century onwards (in particular after the 1980s), Buddhist oral teachings by the 14th Dalai Lama and other Tibetan lamas were recorded, transcribed and published as (electronic) books, a selection of which were added to PACTib as well. The Modern Spoken Tibetan had by that time already started to diverge significantly from the the Classical Literary language, but transcriptions of oral teachings are often edited to make them more similar to the written standard. In addition to oral teachings, at the end of the 20th century a number of Tibetan novels were published on a variety of topics. From the 21st century, we include collections of Tibetan blog posts and online newspaper articles, as well as the transcribed version of the Spoken Tibetan Corpus consisting of folktales, songs and other fieldwork done in the early 2000s in Tibet (Germano et al., 2017). All subcorpora differ significantly in size, ranging from $\sim 22k$ tokens in Old Tibetan to collections of millions of tokens from the BDRC as well as the translated Buddhist canon. For our present purposes, we aimed to annotate everything that was available in digital form and could be dated. In future work, when more studies of the materials become available, more careful selections can be made to create a more balanced annotated corpus suited for specific research questions.

### 2.1 Verse vs Prose

Because metadata for all of our subcorpora is extremely limited or non-existent, it is impossible to distinguish between verse and prose texts.[3] Automatic detection of verse is often done based on phonetic structure and rhyme (cf. Kesarwani (2018)). Since these features do not necessarily characterise Tibetan verse, we searched for other features. In Tibetan verse, the end of a line is always indicated by a ། *shad* marker. In prose texts, these *shad* markers can function as the equivalent of commas in enumerations, but are also used as semi-colons, colons or at the end of sentences. Since Tibetan verse lines are short (generally nine syllables at most), poetic texts have a much higher number of *shad* markers than prose equivalents of comparable length. This ratio of *shad* could thus be used as a very rough indicator of whether we are dealing with verse or prose.

For each text in our corpus we thus calculated the '*shad*-index' (the ratio of *shads* and overall tokens) and found a variety of 4.2-15.3: the higher the *shad*-index, the more likely it is that the text contains a large amount of verse. We verified the range with a known poetic text with verse lines of nine syllables (i.e. a long verse line in Tibetan, thus indicating a low boundary of the *shad*-index). This poetic text (Karu, 1974) has a *shad*-index of 10.41. It therefore seems reasonable to use a *shad*-index of 10.0 as a cut-off point when using the treebank for syntactic

[2]More Bon texts are available and some of those are already digitised: as soon as they become publicly available, we will incorporate them in PACTib.

[3]Note that in Table 1 we provisionally mark the topics or general text genres when they are commonly known; more specific information on verse vs prose, however, does not exist for most texts in our corpus.

queries in particular that are likely to be influenced by poetic styles. We briefly touch on this in Section 5.2.

A further indication that this cut-off point is on the right track is provided by the results of online news articles and blog posts from the 21st century, which we know are not focused on poetry. They have a *shad*-index of 5.08 and 4.78 respectively. Finally, it is important to note that the Old Tibetan *Annals* and *Chronicle* have a *shad*-index of 12.48 and 12.07 respectively, which would thus place them on the poetic side of the divide according to our calculations. However, the Old Tibetan language can be characterised by a number of features that distinguish it from Classical Tibetan. For instance, the texts are known to be formulaic in nature (Takeuchi, 2011) and in addition there are specific features of the punctuation that drive up the number of *shad*s per token (e.g. །།། ་།།) resulting in a higher *shad*-index than we would normally expect for known prose texts in Classical and Present-Day Tibetan.[4]

## 3 Linguistic Annotation

The linguistic annotation of PACTib consists of tokenisation, sentence segmentation, part-of-speech tags and syntactic phrase structure labels building for a constituency treebank on recent work by Meelen and Hill (2017) and Faggionato and Meelen (2019). We optimised their methods after an error analysis and for the purposes of this paper, focused mainly on creating meaningful sentence segmentation.

### 3.1 Tokenisation and sentence segmentation

The Tibetan script has no markers to indicate word and sentence boundaries. Alongside morphosyntactic information, the linguistic annotation for PACTib therefore necessarily includes tokenisation and sentence segmentation as it can have consequences for any subsequent NLP tasks like part-of-speech (POS) tagging or Named Entity Recognition (NER) as well as for diachronic linguistic studies of the corpus. Tokenisation of PACTib was done using Meelen and Hill (2017)'s method combining memory-based syllable tagging and rule-based recombination of syllables into words. Clitics and case markers were considered separate tokens to reduce the overall number of different morphosyntactic tags. Sentence segmentation in the most recent version of the ACTib (Meelen et al., 2017) was purely done automatically, with utterance boundaries added after the Tibetan punctuation marker ། *shad* or ། ། *double shad*. The single *shad* in particular, however, is often more like the equivalent of a comma in English, as it is used in enumerations and subordinate clauses as well. When doing syntactic research on the clause or sentence level in particular, these forced sentence fragments are often too short to yield meaningful data. For this parsed version of ACTib, we therefore aimed to optimise the segmentation of sentences in a linguistically informed way through a series of rule-based replacements combining sentence fragments to fully grammatical sentences and splitting up combinations of what we would consider main clauses.

As a rigid head-final language, Tibetan exhibits object-verb (OV) order (DeLancey, 2003a) and verbs therefore always appear at the very end of the clause or sentence. Although Tibetan verbs exhibit no person-number agreement affixes, overt tense/aspect/mood (TAM) markers are attached to the right of verbal stems. In addition, Tibetan verbal forms can be nominalised with a variety of nominalisation suffixes. Nominalised verbs (with their arguments) do not function as the main verb of the sentence and were therefore, unlike their verbal 'conjugated' counterparts not used to identify sentence boundaries, as shown in example (1), where the nominalised verb *bkru* 'wash' (in bold) is not the matrix verb, but modifying the noun *dkaryol* 'cup' instead:

(1) དཀར་ཡོལ་བཀྲ་ཡག་དེ་ཚོ་གར་ཡོད་རེད་
*[NP karyɔl* ***bkru*** *yag detsho] gapar yod red*
cup wash NOM these where EXIST.COP

[4] See Dotson and Helman-Ważny (2016, 82-85) for a detailed overview of punctuation and the use of *shad* and other markers in Early Tibetan documents. In future work we will refine our methods for the *shad*-calculation to be able to deal with specific orthographic features that lead to aberrant *shad* counts like these.

‘Where are the cups to be washed?’ (i.e. that need washing) (Tournadre and Dorje, 2003, 178)

The verb stem *bkru* in (1) would receive a verbal part-of-speech tag, but as it is followed by the nominaliser *yag*. In addition to conjugated verbs at the end of matrix clauses, Tibetan can exhibit sentence-final particles འོ་ དོ་ ཏོ་ ནོ་ བོ་ མོ་ སོ་ ངོ་ རོ་ *’o do to no bo mo so ngo ro* that indicate the end of the sentence. Finally, for the purpose of correcting parsed structures and a range of syntactic research it is more convenient to split coordinate main clauses into two separate sentences (Meelen and Willis, 2020). Therefore, conjugated verbs that are followed by the conjunction དང་ *dang*, tagged as an associative converb (`cv.ass`) are also followed by a sentence boundary. Sentence boundaries were therefore inserted according to the following set of sequential rules:

1. conjugated verbs + `cv.ass` + *shad*[5]
2. conjugated verbs + (final particle) + *shad*
3. final particles + (*shad*)

This ordered set of rules yields sentence boundaries that form a major improvement on the automatically added utterance boundaries after every *shad*, because *shad* is also used as the equivalent of a comma or semi-colons, resulting in each item of enumerations etc. (of which there are generally many in Buddhist texts) ending up as separate sentences that are not well-suited for morphosyntactic research.

### 3.2 POS tagging and Parsing

POS tagging was initially done with the memory-based method developed by Meelen and Hill (2017), but extended with a number of further rule-based corrections (e.g. erroneously tagged དང་ *dang* ‘and, (together) with’ > `case.ass` ‘associative case marker’, since in the context directly following nouns, it can never be anything else). Syntactic phrase-structural information was added using the rule-based regular expression parser developed by Faggionato and Meelen (2019) that combines Tibetan POS tags into phrases using an extended form of the NLTK’s regular expression chunkparser. This form of constituency parsing was chosen to facilitate comparative historical syntactic research on phrase structure in the UPenn historical treebank tradition. However, unlike the UPenn historical corpora, we deliberately chose not to add empty categories of any kind, to make PACTib more theory-neutral and because manual correction (which is always necessary as automatic insertion and annotation of empty categories is very prone to error) of such a large corpus is impossible. Another reason to create semi-hierarchical structures only and avoid empty categories for the present corpus is that the resulting bracketed structures can easily be converted to a dependency treebank format in combination with our highly detailed morphosyntactic tag set. Finally, attempts to develop automatically parsed dependency treebanks for Tibetan are already being undertaken by the researchers at SOAS, University of London, in the context of the ‘Lexicography in Motion’ project (Faggionato and Garrett, 2019) so a constituency-based treebank fills this gap in the literature. Example (3) shows the parsed result of a simple transitive clause like (2):

(2) ངས་ཁ་ལག་བཟས་པ་ཡིན།

*[NP ngas] [NP kha lag] [VP bzas pa yin]*
I.ERG food ate.PAST

‘I have eaten the food’ (Tournadre and Dorje, 2003, 165)

[5]For *shad* here, we mean any variety of Tibetan punctuation marker that conveys a function similar to the single *shad*. Depending on the text type or genre, variants like ། *gnyis shad* or “double” *shad*, ཿ *gter tsheg* or ། *tsheg shad* are used as the equivalent of commas, semi-colons, colons or full stops, just like regular *shad*.

(3) (S (NP དངོ/p.prop )
(NP ཁ་ཡག/n.mass )
(VP བཟས་པ་ཡིན/v.past)
(PUNC །/punc ))

Hierarchical structures, e.g. noun phrases within postpositional phrases are also automatically captured:

(4) བོད་ལ་གནམ་གྲུ་ཡོད་རེད།
*[PP [NP bod] la ] [NP gnam gru] [VP yod red]*
Tibet in aeroplanes EXIST.COP
'There are aeroplanes in Tibet.' (Tournadre and Dorje, 2003, 121)

(5) (S (PP (NP བོད/n.prop ) ལ/case.all ))
(NP གནམ་གྲུ/n.count
(VP ཡོད་རེད།/v.pres ) (PUNC །/punc ))

Since the rule-based parser and memory-based taggers were originally developed for Old and Classical Tibetan texts respectively, they are not always optimally suited for the Present-Day Spoken Tibetan language, which has evolved in a number of ways. Present-Day Literary Tibetan (or any form of the present-day written language) still strongly resembles Classical Tibetan (see also Section 5). Present-Day Spoken Tibetan nominalisation markers like ཡག *yag* or གར *gar* that do not exist in Classical Tibetan receive a special POS tag `nom`, which only exists in transcribed oral texts in Present-Day Tibetan. Since evidential, egophoric and epistemic verbal endings in Present-Day Tibetan have evolved from homophonous verbs and TAM markers in Classical Tibetan we chose to retain the conservative morphosyntactic annotation for those to facilitate research on diachronic changes in this aspect of the grammar.

Finally, it is important to note that Present-Day Tibetan contains a range of modern vocabulary items that are not found in the Old and Classical Tibetan training data. This goes for a number of modern verbs, e.g. ཕབ་ལེན *phab len* 'to download'. Most of these verbal forms, however, are based on combined verbs or light-verb constructions that already exist in Classical Tibetan and thus provide no real issue when conservative noun or verb tags are used, e.g. *phab len* 'download' < *phab* 'to bring down' + *len* 'to take', *kha par btang* 'to make a phone call (to)' < *kha par* 'phone' + *btang* 'to send'. Other new vocabulary, mainly from after the industrial and technological revolutions, mostly consists of nouns. Since count nouns (tagged `n.count`) are by far the most frequently-occurring tags, the memory-based tagger (and the neural tagger developed by Faggionato and Meelen (2019)) mainly assign this `n.count` tag to unknown words in the right context, these new vocabulary items pose no significant problem in Present-Day Spoken Tibetan texts.

## 4 Retrieving and Adding the Metadata

The PACTib is not only unique because of its size and scope, but also because it is the only Tibetan corpus with meaningful metadata linked to every sentence. As discussed in Section 2.1, there is in fact hardly any metadata available for any of the digitised texts that are available. Present-day oral teachings can of course be linked to known lamas and the connections can sometimes be made for well-studied historical texts, such as the works by Mipham in the 19th century and the *gZer mig*. The *Annals* and *Chronicle* have been the main focus of study for scholars of Old Tibetan as well, but they still disagree about the date of origin (ranging from the 9-11th century). Since our current main objective is to create an annotated diachronic corpus suitable for morphosyntactic research, our first aim is to attempt to link *all* the digitised materials in our subcorpora to meaningful dates of origin. Although the e-texts from the BDRC collection did not come with any readily available metadata, it is possible to get an idea about the date of origin because information about the author or a patron of a text (when this is

available) is linked to the textIDs of e-texts in the BDRC database, which contains over 21,000 e-texts in total. For many of these authors and patrons, there is furthermore information about either their date of birth, date of death or both. Although this does not give us an exact date of origin for each text, it does provide us with a date range, which can be used to derive an approximate date of origin. We therefore extracted the date range of the life of an author or patron associated to the e-texts for which this was available (a total number of just over 5,000 e-texts, which is about a quarter of the total BDRC collection) from the BDRC's database, using SPARQL queries on:

- the date when the text was composed (rarely known)
- the birth/death date of the main author or patron

In this way, only texts where either the composition date or the birth/death date of the author was available, were added to our corpus. The BDRC has a rather large database, including 18,000 persons (authors, editors, important historical figures) and 40,000 books, historically focused on the Tibetan cultural area. It has recently moved to LOD (Linked Open Data) and is now able to aggregate results from datasets from partner organisations, such as the Sakya Research Centre or the Treasury of Lives, both of which also contain information about Tibetan authors. Finally, we were able to extract additional information regarding the topic of some of the texts. We could thus partially address the issues concerning the lack of metadata by extracting as much information as possible from a range of available resources, combining it in one place and making it accessible (see our annotated corpus and metadata files deposited on Zenodo through the link on our ACTib GitHub repository where all code and queries can be found as well). As the number of partner organisations willing to share their data with the BDRC grows, more and more data will be available on each author, thus allowing more and more e-texts to be added to future versions of this corpus. These dates were made an integral part of the SentenceIDs that were automatically added to all sentences in the treebank. Making dates/date ranges available through the SentenceIDs means the treebank can be queried in any way and results can be easily organised by date, without relying on any further resources. In the next Section we demonstrate this with two short case studies.

## 5 Tracing Diachronic Stability & Change

To illustrate potential uses of PACTib in this section we present two short case studies of diachronic morphosyntactic research questions that can be investigated with our treebank. Both case studies are based on observations by Tournadre and Dorje (2003) in their section on differences between Classical/Literary Tibetan and Present-Day Spoken Tibetan.

### 5.1 Oblique Case Markers

Our first case study is a change in the use of case marking particles. Old and Classical Tibetan exhibit a wide range of oblique case markers or postpositions, that vary in form due to their specific phonological contexts (DeLancey, 2003a). Each of these case markers are split off from the preceding words and tagged as `case.all` for 'allative/dative',[6] `case.loc` for 'locative', etc. As Tournadre and Dorje (2003) note, from the outset dative/allative *la*, locative *na* and terminative *du* (and their phonological variants *-r, ru, su, tu*) could function as the locative indicating a specific place (without movement), as shown in example (6):

(6) བོད་དུ་ བོད་ལ་ བོད་ན་
*bod* ***du;*** *bod* ***la;*** *bod* ***na***
Tibet TER Tibet ALL Tibet LOC
'in Tibet' (Tournadre and Dorje, 2003, 413)

[6] We follow Hill (2007) here calling Tibetan ལ་ *la* the allative marker although it has a range of other functions, e.g. dative, as well, which is why some refer to this as the dative marker (cf. Tournadre and Dorje (2003)).

In Present-Day Spoken Tibetan, in particular in Lhasa Tibetan, the dative/allative case marker *la* has taking over the functions of more and more other oblique case markers, leaving the locative, terminative, etc. as 'relict' forms such as adverbs and complex postpositions (see Section 5.2) only (DeLancey, 2003b, 275), as shown in example (7):

(7) སྦྱིར་དུ་ ལྷག་པར་དུ་
*sger* ***du;*** *lhag par* ***du***
private TER specially TER
'privately, personally; especially' (DeLancey, 2003b, 275)

If we query our treebank looking for postpositional phrases with allative/dative markers as opposed to other oblique cases, we can clearly see a rise of the use of allative *la* at the expense of terminative *du* in particular, as shown in Figure 1. The observations by Tibetan scholars such as DeLancey (2003b) and others that were based on the manual comparison of a small number of Tibetan texts from different time periods were definitely on the right track: in the modern spoken UVA subcorpus in particular we can see this change. The corpus also show a slight rise in dative/allative markers in 21st-century books, but this does not hold for online news articles and blogposts from the same period. This indicates that although the written language has evolved, it is still very far removed from the modern spoken language represented here by the Present-Day UVA subcorpus. Finally, it is actually quite remarkable how stable the distribution of oblique markers is across 11 centuries. From the 11th century onwards, terminative markers form the clear majority, which is not surprising as they have a very wide range of other functions besides the locative of place. Functions of elative, ablative and locative markers are much more restricted, which is clearly reflected in the data.

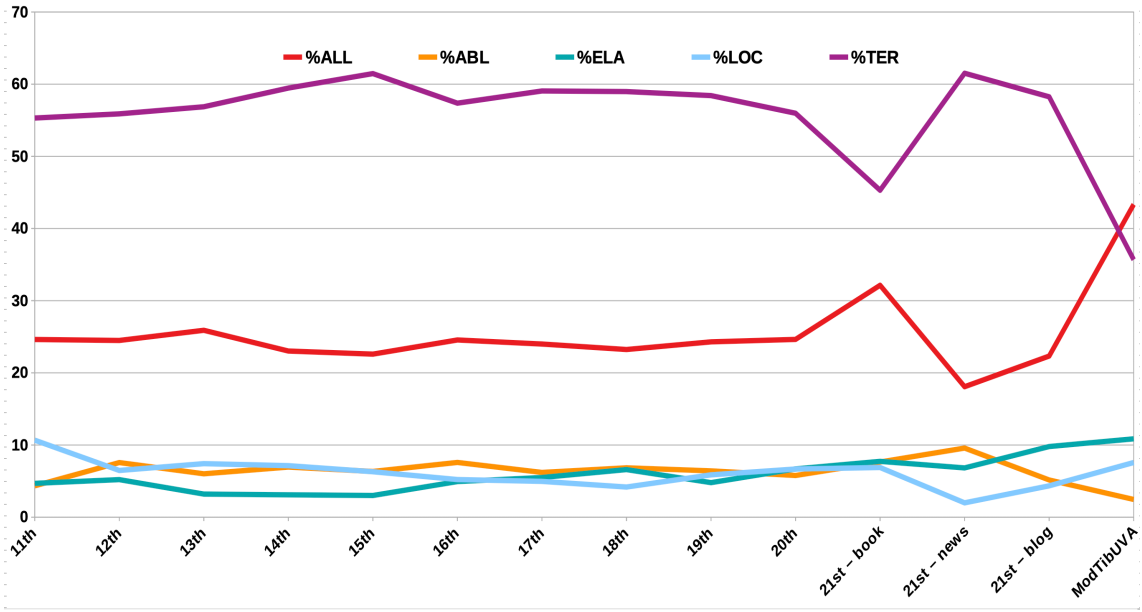


Figure 1: Ratio of oblique case markers from 11th-21st centuries.

## 5.2 Complex Postpositions

Our second case study concerns the syntax of complex postpositions that are tagged as 'relator nouns' (`n.rel`) in our treebank. These postpositions are originally lexical nouns that through a process of grammaticalisation have changed into functional items in combination with a noun phrase followed by a genitive case marker. The postposition itself can furthermore be followed by an oblique case marker (allative/dative, ablative, elative, locative or terminative). An example with the postposition *nang*, originally a noun meaning 'inside' but now part of the complex postposition preceded by a genitive and followed by a terminative case marker, is shown in (8):

(8) བོད་ཀྱི་ནང་དུ་ཡི་གེ་འབྲི་སྟངས་བཞི་ཡོད་པ་རེད་

*bod (kyi) **nang** (du) yige 'bri stangs bzhi yod pa red*
Tibet GEN inside TER letter writing styles four EXIST.COP

'There are four styles of writing in Tibet.' (Tournadre and Dorje, 2003, 410)

Tournadre and Dorje (2003, 410) observe that in Classical/Literary Tibetan the preceding genitive case marker and the following oblique case markers are optional, whereas in Present-Day Spoken Tibetan these case markers cannot be omitted. Note that for poetic texts with verse lines forced into a predetermined number of syllables (often 5, 7 or 9), deliberate use *or* omission of the genitive marker to make up the right amount of syllables can be expected. Evidence for this particular construction in which the use of the genitive marker is believed to be optional in Old and Classical Tibetan could in theory thus go either way. A complete study of this goes beyond the scope of our present paper, but in future work, we will use the *shad*-index we established in Section 2.1 to test various hypotheses along these lines. If this is a gradual process of change, we would expect an increase in the use of genitive markers at the end of the Classical Tibetan period leading to a ratio of almost 100% genitive case markers in the 21st century, in particular in the spoken UVA subcorpus. Figure 2 shows the results of our complex postpositions with and without preceding genitive case markers. Percentages of the use of preceding genitives with postpositions are split up into different categories determined by the following oblique case markers (allative/dative, ablative, elative, locative and terminative) or '%gen-N', for the final option without final case marker.

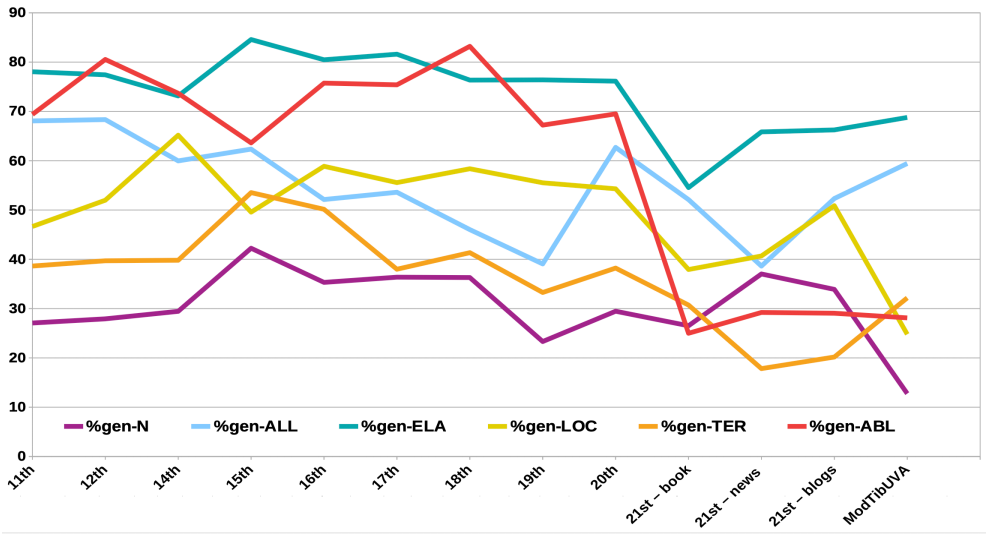


Figure 2: % of preceding genitive case markers in complex postpositions.

A initial interesting observation concerning this variable in our 166*m*-token corpus is again one of remarkable stability as we have seen with the oblique case markers above: the use of genitive case markers in this construction remains relatively stable between the 11th and 20th centuries.[7] Confirming Tournadre & Dorje's observation, the genitive marker was indeed often omitted in these constructions. However, we can observe a clear distinction between those complex postpositions followed by an ablative, allative or elative, where from early on genitive markers were use around 70-80% of the time, whereas numbers for complex postpositions with locatives and terminatives (or without following postpositions for that matter) are much lower.

As we expect changes to occur from the 20th century onwards, we again show the different sources in the 21st century in further detail by book, online news articles, blog posts and the transcription of the Present-Day Tibetan spoken UVA corpus. Interestingly, the use of genitives

[7]Note that the 13th century was omitted here because the lack of data from this period distorts the results of queries for lower-frequency constructions like these complex postpositions.

markers appears to decline at first compared to the previous centuries, but in the blog posts and in particular in the spoken UVA subcorpus, the use of genitive case markers is rising again.[8] The genitive is still not found 100% of the time, however, and numbers for the combinations with locative and ablative case markers are particularly low. The main reason for this is due to scarcity of data, which we will discuss in the final Section.

### 5.3 Discussion and Limitation of this first version of PACTib

In this final Section we discuss the results of our case studies in light of potential issues and limitations with this first version of PACTib. First, as we already noted, this is not a balanced corpus, but instead a collection of all the digitised Tibetan materials that were available to us. If a research query depends on a well-balanced corpus, it would be good to try and make a selection of selects from the PACTib to achieve this goal. As more and more Tibetan texts are being digitised these days, we expect the PACTib can soon be extended and gaps in time and genre can be filled. From the historical period, it would be good to have more material from the 11th and 13th centuries. From the modern, it would be good to collect more spoken material from a range of Present-Day Tibetan varieties, as the written and spoken language clearly differs and even 21st-century blog posts do not necessarily reflect the language as it is spoken today. As an example from our case study, the low number of locative and ablative case markers in the modern spoken subcorpus distorts the ratios. There are, for instance, only 32 cases of complex postpositions with ablatives overall; 9 of which have the genitive, a ratio of 28.12%. If we look at the numbers for the allative, elative and terminative on the other hand, we get hundreds of examples, therefore showing more robust patterns along the lines of what we expect. Scarcity of data is also an issue for the 21st-century book, which is with only 128,716 tokens, significantly shorter than the contemporary subcorpora containing news articles and blogposts (over 3 million tokens each).

Apart from data scarcity for certain constructions in specific subcorpora, this ablative case marker example illustrates a final limitation of this first version of the PACTib, namely, the possibilities of errors in the annotation. Tibetan ལྟ་ *lta* , for example, is often tagged as `n.rel` in the training data, because it can indeed have that function with the meaning 'like N' (following the noun N and potential genitive). However, *lta* has a range of other meanings as well and occurs in various phrases and expressions in which its special signification (derived from the verb 'to see') is no longer clearly discernible (Jäschke, 1987, s.v. ལྟ་). In the spoken UVA subcorpus, for example, we find a number of examples with དེ་ལྟ་ *da lta* where *lta* is still tagged as `n.rel`, even though in this sequence it actually means 'now...' and a preceding genitive would be impossible. Some results with the ablative case marker *las* in the spoken subcorpus are in fact cases of tagging errors: the sequence དེ་ལྟ་ལས་བཟོ་ *da lta las bzo*, for example, was tagged as a complex postpositional phrase with ablative *las*, and counted as a result without a genitive marker. In fact, the segmenter here failed to segment the disyllabic noun ལས་བཟོ་ *las bzo* 'worker' properly and instead identified *las* as an ablative case marker that was part of a complex postpositional phrase. Because our corpus was automatically annotated with tools developed for Classical Tibetan, errors in annotation can always occur and affect the results. With frequent or less complex queries like our case study on oblique case markers in Section 5.1, this is not problematic as despite their ambiguous nature, the Precision and Recall of simple case markers following nouns is very high (Meelen and Hill (2017, 83-85) report an F-score of 0.99 for `case.term` and `case.all` and 0.98 for `case.abl`). With more complex or less frequent constructions more care should be taken. When segmentation has gone wrong in a sequence of syllables that are all highly ambiguous, as is the case of the above example in the context of multifunctional *da* and *lta* followed by the wrongly segmented single syllable *las*, this can affect the results. In this particular case this was exacerbated by the fact that there are relatively few

[8] This is not the case for the %gen-N context without oblique case markers, which is probably due to a change in the use of oblique case markers in general and specific postpositions with changed meaning in Spoken Tibetan, an investigation of which goes beyond the scope of the present paper.

instances of the ablative in Present-Day Spoken Tibetan to begin with. With corrections in a post-processing stage, as suggested by Meelen et al. (forthcoming), some of these issues can be addressed. However, for Present-Day Spoken Tibetan, it would ultimately be best to train a segmenter and tagger on contemporary spoken data, instead of relying on those developed for Classical Tibetan.

## 6 Conclusion & Future Work

In this short paper we present the first historical Tibetan treebank: the PArsed Corpus of Tibetan. PACTib is a linguistically annotated corpus of $> 166m$ tokens with dates ranging from the 11th to the 21st century. This corpus brings together all digitised historical materials that were available to us and for which at least a rough date of origin could be defined. Dates of origin derived from information about authors/patrons associated with the texts were extracted from the BDRC's database, which is partially fed with information through Linked Open Data protocols and agreements with partner organisations. This information was then systematically added not just to PACTib's metadata file, but also to all SentenceIDs so that results from corpus queries can be easily organised by date. The linguistic annotation consists of word and sentence segmentation, POS tags and constituency-based phrase structure. Our new method of sentence segmentation based on linguistic features means that parsing can be done efficiently and the resulting treebank facilitates any kind of syntactic research of longer and more complex sentences as well. In addition, the metadata for our treebank contains information about the number of tokens as well as the topic of the text (when available). Finally, we proposed the '*shad*-index', the ratio of the Tibetan punctuation marker *shad* and the total number of tokens, that indicates the likelihood of the text containing large amounts of verse. Because there is no information available on the genre of most of these texts, nor is there another way to automatically distinguish prose from poetry, which would be particularly useful for syntactic research. Our first attempt at calculating the *shad*-index of a text could be refined by critically examining more of our source materials, making sure that ornamental sequences of punctuation markers like *shad* such as those in the Old Tibetan texts are not skewing the results, but a first test with some known verse vs prose texts already yields promising results.

We finally presented two short case studies to illustrate how PACTib can be used for morphosyntactic research and to test the limits of the current version. With case studies on oblique case markers and complex postpositions we demonstrate PACTib can be a useful tool to test hypotheses on diachronic morphosyntactic developments. One interesting conclusion from both is that the Tibetan language has remained remarkably stable for over a thousand years in these two aspects of grammar. The main limitations are currently the lack of (balanced) data (especially for the 11th and 13th centuries, as well as the present-day spoken subcorpus) and certain issues with errors in the automatic annotation of ambiguous forms. We addressed some of the latter in forthcoming work (Meelen et al., forthcoming), but acknowledge that in order to really improve the annotation of Present-Day Spoken Tibetan, it would be best to train taggers on data from manually corrected Present-Day Spoken corpora once they become available.

## Acknowledgements

The authors gratefully acknowledge Meelen's British Academy Postdoctoral Fellowship pf170063 grant and the ERC Advanced Grant 'Open Philology' (No 741884) for partially funding the research time of the authors that led to this paper and three anonymous reviewers for their helpful comments. In addition, the authors gratefully acknowledge the following individuals who made electronic versions of Tibetan texts available: Robbie Barnett (for the collection of Tibetan online newspaper articles), Gregory Forgues (for the collection of Mipham works), Edward Garrett (for the collection of Present-day Tibetan blog posts), Charles Ramble (for the *gZer mig* and a section of a 20th-century novel) and Ngawang Trinley (for support with the selection of a number of oral teachings and for his work on the *eKangyur* and *eTengyur*).

## References


Scott DeLancey. 2003a. Classical Tibetan. In Graham Thurgood and Randy J. LaPolla, editors, *The Sino-Tibetan Languages*, volume 3, pages 255–269. London/New York: Routledge.

Scott DeLancey. 2003b. Lhasa Tibetan. In Graham Thurgood and Randy J. LaPolla, editors, *The Sino-Tibetan Languages*, volume 3, pages 270–288. London/New York: Routledge.

Brandon Dotson and Agnieszka Helman-Ważny. 2016. *Codicology, Paleography, and Orthography of Early Tibetan Documents: Methods and a Case Study.* Arbeitskreis für Tibetische und Buddhistische Studien Universität Wien.

Christian Faggionato and Edward Garrett. 2019. Constraint Grammars for Tibetan Language Processing. In *NEALT Proceedings Series 33:3*, pages 12–16.

Christian Faggionato and Marieke Meelen. 2019. Developing the Old Tibetan treebank. In Nikolova Temnikova Angelova, Mitkov, editor, *Proceedings of Recent Advances in Natural Language Processing*, pages 304–312. Varna: Incoma.

Edward Garrett, Nathan W Hill, and Abel Zadoks. 2014. A rule-based part-of-speech tagger for Classical Tibetan. *Himalayan Linguistics*, 13(2).

David Germano, Edward Garrett, and Stephen Weinberger. 2017. Uva tibetan spoken corpus, June.

Nathan W Hill. 2007. Aspirated and unaspirated voiceless consonants in Old Tibetan. *Languages and Linguistics*, 8(2):471–493.

Heinrich August Jäschke. 1987. *A Tibetan-English Dictionary: with special reference to the prevailing dialects, to which is added an English-Tibetan vocabulary.* London: Routledge & Kegan Paul Ltd.

Grub-dbang Karu. 1974. *The Autobiography of dKar-ru Grub-dbang bsTan-'dzin rin-chen. (dPal snya chen rig 'dzin mchog gi rnam sprul bāi'u ldong btsun grub pa'i dbang phyug bstan 'dzin rin chen rgyal mtshan bde chen snying po can gyi rnam par thar pa rmad 'byung yon tan yid bzhin nor bu' i gter).* Dolanji: Tibetan Bonpo Monastic Centre.

Vaibhav Kesarwani. 2018. *Automatic Poetry Classification Using Natural Language Processing.* Ph.D. thesis, Université d'Ottawa/University of Ottawa.

Huidan Liu, Minghua Nuo, Longlong Ma, Jian Wu, and Yeping He. 2011. Tibetan word segmentation as syllable tagging using conditional random field. In *Proceedings of the 25th Pacific Asia conference on language, information and computation*, pages 168–177.

Marieke Meelen and Nathan Hill. 2017. Segmenting and POS tagging Classical Tibetan using a memory-based tagger. *Himalayan Linguistics*, 16(2).

Marieke Meelen and David Willis. 2020. Towards a Historical Treebank of Middle and Early Modern Welsh, Part I: Workflow and POS Tagging. *Journal of Celtic Linguistics*, 22(1):125–154.

Marieke Meelen, Nathan W. Hill, and Christopher Handy. 2017. The Annotated Corpus of Classical Tibetan (ACTib), Part I - Segmented version, based on the BDRC digitised text collection, tagged with the Memory-Based Tagger from TiMBL, July.

Marieke Meelen, Élie Roux, and Nathan Hill. forthcoming. Optimisation of the largest annotated Tibetan corpus combining rule-based, memory-based deep-learning methods. *Transactions on Asian and Low-Resource Language Information Processing.*

David L Snellgrove. 1967. The nine ways of Bon: excerpts from " gZi-brjid". *London oriental series.*

Tsuguhito Takeuchi. 2011. Formation and transformation of old tibetan. In T. Takeuchi and N. Hayashi, editors, *Historical Development of the Tibetan Languages. Proceedings of the Workshop B of the 17th Himalayan Languages Symposium, Kobe, 6th–9th September*, pages 3–17.

Nicolas Tournadre and Sangda Dorje. 2003. *Manual of standard Tibetan: Language and civilization: Introduction to standard Tibetan (spoken and written) followed by an appendix on classical literary Tibetan.* Boston: Snow Lion Publications.