Developing the Old Tibetan Treebank

Christian Faggionato
SOAS, University of London
cf36@soas.ac.uk

Marieke Meelen
University of Cambridge
mm986@cam.ac.uk

Abstract

This paper presents a full procedure for the development of a segmented, POS-tagged and chunk-parsed corpus of Old Tibetan. As an extremely low-resource language, Old Tibetan poses non-trivial problems in every step towards the development of a searchable treebank. We demonstrate, however, that a carefully developed, semi-supervised method of optimising and extending existing tools for Classical Tibetan, as well as creating specific ones for Old Tibetan, can address these issues. We thus also present the very first Tibetan Treebank in a variety of formats to facilitate research in the fields of NLP, historical linguistics and Tibetan Studies.

1 Introduction

In historical linguistics, there are currently two types of morpho-syntactically annotated corpora or 'Treebanks', one based on constituency parses, the other on dependency parses. The former includes pioneering work in the Penn-Helsinki tradition, resulting in the Old and Middle English (Taylor and Kroch, 1994), Icelandic (IcePaHC) (Rögnvaldsson et al., 2012) and Portuguese (Tycho Brahe) (Galves, 2018) treebanks. The latter represents syntactic structure in the form of dependencies and is often used for applied NLP tasks and early Indo-European languages, e.g. the PROIEL Treebank family (Eckhoff et al., 2018). In this paper we present a constituency-based treebank for an under-resourced language: Old Tibetan (see Green et al. (2012) and El-Haj et al. (2015) for similar examples of creating under-resourced treebanks).

Old Tibetan is an extremely under-resourced and under-researched language from an NLP point of view. We chose to focus our attention on the Old Tibetan corpus (7-11th c.) since it consists of a small collection of documents compared to the vast amounts of translated and original Classical Tibetan texts. Nonetheless, the Old Tibetan corpus is still heterogeneous enough to represent natural language. The majority of Old Tibetan texts (known to date; new inscriptions and texts are still being discovered) has now been digitised in one way or another (images, OCR and/or transcribed) and annotating this data is fundamental for the understanding of diachronic and synchronic issues in Tibeto-Burman languages. As the first attempt to create a Tibetan Treebank, developing a segmented, POS-tagged and chunk-parsed corpus of Old Tibetan provides new opportunities in Tibetan scholarly history, literature and linguistics.

Old Tibetan was the language spoken in the Yarlung Valley from where the Tibetan empire started its initial expansion. Writing was mainly introduced to facilitate administrative tasks, and the earliest Old Tibetan texts represent the most detailed sources for the history of early Tibet (Hill, 2010). The earliest currently available, securely datable Old Tibetan document dates to ca. 763 CE. However, the digital resources for Old Tibetan are inadequate (problematic transcriptions, transliterations and no digitised secondary resources such as dictionaries, etc.).

The core of the Old Tibetan corpus is available as plain e-texts (without segmentation or any kind of annotation) on the Old Tibetan Documents Online (OTDO) website.[1] We first focus on the Old Tibetan *Annals* (5.9k tokens)

[1] http://otdo.aa.tufs.ac.jp/

Proceedings of Recent Advances in Natural Language Processing, pages 304–312,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_035

and Old Tibetan *Chronicles*, since these are the best sources that we have at our disposal in terms of length and linguistic variety (many other Old Tibetan text are short inscriptions or more fragmentary). The Old Tibetan *Annals* are Tibet's earliest extant history. The Old Tibetan *Chronicle*, written in the early 9th century, is more narrative and includes historical accounts and songs related to the Yarlung dynasty and the Tibetan empire.

In this paper we present our annotation procedure that addresses all issues of pre-processing, segmentation, POS tagging and parsing in detail. Our semi-supervised method, resulting in the first Old Tibetan Treebank, can furthermore serve as an example of how to overcome challenges of low-resource and under-researched languages in general.

2 The Annotation Procedure

Since Old Tibetan is an extremely under-resourced language the procedure to develop an annotated corpus needs to be developed with great care. Additional steps are necessary at each of the normal stages, from pre-processing to POS tagging and parsing and finally post-processing. In the pre-processing stage, for example, the normalisation is not a trivial task because of a range of issues with the Tibetan script and the way it is digitised, in either Unicode or a variety of transliteration formats. In addition to solving these script issues, our core solution is to transform our Old Tibetan texts through a 'conversion/normalisation process' with a Constraint Grammar (Cg3) to a form of Tibetan that is closer to Classical Tibetan, for which at least some NLP tools are available.

Before we can move on to the annotation stage, we need to solve a further non-trivial issue of finding word and sentence boundaries. Since there are no Gold Standards or training data available for Old Tibetan, we resort to the little material and tools available for Classical Tibetan and then do a rigorous error analysis checking specific Old Tibetan features that we know differ from Classical Tibetan. Our annotation method is thus supervised in various ways to overcome the obstacles building a treebank of an extremely low-resourced language like Old Tibetan.

3 Pre-Processing

The Old Tibetan texts we work with to start this corpus are already transcribed from the original manuscripts or digitised images. For the present paper we thus only address the issues concerning encoding of transcriptions and transliterations and the issues of tokenising a language without word or sentence boundaries.

3.1 Transliteration Issues

One of the first challenges we encountered in creating the Old Tibetan corpus was the conversion from Tibetan Unicode script (see Hill 2012) to the Wylie transliteration system. There are few reliable tools available and in addition, we have to take the peculiar orthographic features of Old Tibetan into consideration. The Tibetan Unicode script for the Old Tibetan documents was obtained from a modified version of the Wylie transliteration system that is used for the Old Tibetan Documents Online (OTDO) website, through the BDRC conversion tool.[2]

However, this tool only partially addresses the issue, because we also want to transform Old Tibetan into a form of Tibetan that looks more similar to Classical Tibetan in terms of orthography. Therefore, the Wylie transliteration used by the OTDO website had to be modified. As an example the reverse 'i' vowel mark, ◌ྀ - called *gigu* - is transliterated with 'I' on the OTDO website. We substituted 'I' with 'i', which is the standard Wylie transliteration for this character, as shown in (1):

(1) *rgyal po'I* > *rgyal po'i* 'of the king'

3.2 Normalisation

The Old Tibetan script furthermore presents a set of features that need to be 'normalised' or converted to a form that looks like Classical Tibetan. We therefore created a set of rules translated into the Constraint Grammar (Cg3) formalism. Most of the Cg3 rules used to normalise Old Tibetan are simple replacement rules. For example, In Old Tibetan there are many instances of the

[2] https://www.tbrc.org/

above-mentioned reverse *gigu* such as ཀྱི *kyI*. These two forms of *gigu*, ི and ཱི are phonetically indistinguishable and mark no difference in Classical Tibetan. The Cg3 `SUBSTITUTE` rule to normalise the reverse *gigu* is:

```
SUBSTITUTE (``([[\^{ }<]*)\u0F80(.*)"r)
(“$1ི$2v) TARGET (σ)
```

Two additional problems encountered in the normalisation of Old Tibetan are represented by the alternation between aspirated and unaspirated voiceless consonants and the difficulty of splitting merged syllables. This aspiration, however, was probably not phonemic in Old Tibetan (Hill, 2007, 471). Therefore, a set of string replacement rules in the Cg3 formalism was created to normalise and convert these instances to their equivalent reading in Classical Tibetan.

Furthermore in Classical Tibetan, syllables are separated by a punctuation marker called *tsheg*: ་. In Old Tibetan texts, syllable margins are not so clear and syllables are often merged together with the following case marker or converb, e.g. Old Tibetan བཀུམོ *bkumo* > Clas. Tib. བཀུམ་མོ *bkum mo* ‘kill, destroy’:

(2) བཀུམོ > བཀུམ་མོ

These types of merged syllables were also converted to their classical forms, using a set of three regular expressions in the Cg3 formalism through the rule `SPLITCOHORT`. Considering the complexity of the Tibetan syllable, in order to generate the rules, we took the maximum number of its constituents into account (in terms of vowels and consonants) as well as their order.

Generic Rule:

```
([^aeiouI\s]+[aeiouI][^aeiouI\s]*)
([^aeiouI\s'])([aeiouI][^aeiouI\s']*)
> $1$2 $2$3
```

Cg3 rule:

```
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));
```

Through these conversions and normalisations, we could apply existing tools for Classical Tibetan to our Old Tibetan corpus to avoid manually creating our treebank from scratch completely. The full Cg3 grammar is discussed in detail in our forthcoming research.

3.3 Segmenting Sentences

Segmenting sentences is necessary since there are no obvious sentence boundaries in Old Tibetan. The Tibetan scripts does have a punctuation marker that sometimes (but not always) indicates meaningful phrases, a so-called *shad*, ། or double *shad*, ॥. Since without any further annotation, there is no way of knowing where sentences begin or end, we used the single and double *shad* as sentence boundaries and automatically inserted utterance boundaries indicators (`<utt>`) after every instance. This greatly facilitates subsequent annotation tasks that depend on sentence boundaries, such as POS tagging and chunkparsing.

3.4 Tokenisation

The Tibetan script furthermore does not indicate word boundaries. Tokenisation is therefore a tremendous issue, not only for scholars of Tibetan (who often disagree on what the word boundaries should be), but even more so for any Tibetan NLP tasks. The Classical Tibetan script does have a way of indicating syllable boundaries though, by using the abovementioned *tsheg* marker ་, e.g. སྡག་མར transliterated *brag mar* ‘Dagmar’ with spaces between every syllable according to the conventions of the Wylie transliteration.

For Classical Tibetan, Meelen and Hill (2017) addressed this tokenisation issue by recasting it as a classification task with a memory-based tagger (Daelemans et al., 2003) giving ‘beginning’, ‘middle’ or ‘end’ labels to every syllable (automatically split based on the aforementioned *tsheg* and *shad* markers. With our supervised learning method first normalising and then converting our Old Tibetan corpus to a form of Tibetan that is much closer to Classical Tibetan, we were able to

use this existing segmentation tool for Classical Tibetan and extend and modify them after manually correcting part of our Old Tibetan data.

4 POS Tagging

Since there was no Old Tibetan POS-tagged Gold Standard either, here too we started from the Classical Tibetan training data[3] and tagging method developed by Meelen and Hill (2017). We tested a number of ways to get and improve results for the Old Tibetan corpus, e.g. developing a new, reduced tag set, changing scripts (Unicode vs. Wylie) as well as generating new taggers, based on the manually corrected Old Tibetan only and, finally, adding the manually corrected Old Tibetan to the existing Classical Tibetan Gold Standard.

4.1 Small vs Large Tag Set

The tag set used for the Classical Tibetan Gold Standard, developed by Garrett et al. (2014) is with 79 morpho-syntactic tags rather large. This causes major issues for the out-of-vocabulary items, especially for languages without insightful morphological suffixes like Tibetan. For this first attempt of developing an Old Tibetan Treebank, we therefore decided to reduce the amount of tags to a small and simplified version of the standard Universal Dependency POS set, consisting of 15 tags only (De Marneffe et al., 2014). We transformed the existing Classical Tibetan training data, which is our Gold Standard, in the following way: `interj > INTJ`, `punc > PUNCT`, `n.prop > PROPN`, `skt, dunno > X`, `adj, num.ord > ADJ`, `n.v.cop, v.cop, v.cop.neg > AUX`, `n.count, n.mass, n.rel > NOUN`, `num.card, numeral > NUM`, `cl.focus, cv.fin, cv.imp, cv.ques, neg > PART`, `p.indef, p.interrog, p.pers, p.refl > PRON`, `d.dem, d.det, d.emph, d.indef, d.plural, d.tsam > DET`, and, finally, all verb remaining verb forms in all tenses `> VERB`, all remaining converbs `> SCONJ`, all post-positional case markers `> ADP` and all adverbs `> ADV`. A 10-fold cross-validation with the exact same parameter settings of the memory-based tagger[4] on the >318k Classical Tibetan Gold Standard, yielded better results compared to those of the large tag set reported by (Meelen and Hill, 2017) (increase from 95.0% to 96.3% in Global Accuracy; Known Words increased from 96.8% to 97.8%; Unknown Words from 53.4% to 59.7%).

All tags with a very low number of tokens in the out-of-vocabulary set (ranging from $n = 1$-92) have a Precision and Recall close or equal to zero. These items are always very short (one or two characters only), which makes predicting the tag for new items in this category an almost impossible task for the tagger. With the newly trained small tag set tagger, we tagged the Old Tibetan *Annals* and manually corrected the first 3.5k tokens as a start. We then evaluated the tagger again with another 10-fold cross-validation, first on this small Old Tibetan corpus and then again adding this manually corrected Old Tibetan data to the existing Classical Tibetan Gold Standard. This yielded a better Global Accuracy for the combination of Old and Classical Tibetan (96.1%) compared to Old Tibetan alone (92.8%). However, the results for Unknown Words are significantly lower (decrease from 71.1% to 58.5%).

Since these two new Gold Standards differ significantly in size it is impossible to do a fair comparison until we manually correct more Old Tibetan. It is clear, however, that despite our efforts to normalise and convert the Old Tibetan into a form of the language that looks more like Classical Tibetan, it is still making a difference, shown in the lower accuracy (by more than 10%) of unknown words for this combined training data. Without adding the Classical Tibetan training data, however, the vocabulary list that the memory-based tagger builds would simply be too small to get any good results on unseen data. Despite the 10-fold cross-validation, the relatively high scores for the Old Tibetan corpus only are misleading, because of the small size of the corpus. Until we have more manually corrected Old Tibetan data, we therefore proceed with the Classical Tibetan Gold Standard and add an extra stage of error correction, see Section 6.

[3] `http://github.com/tibetan-nlp/soas-corpus/`

[4] These settings for Classical Tibetan are: `-p dwdwfWaw -P psssdwdwdwFawaw -M 1100 -n 5 -% 8 -O+vS -FColumns -G K: -a0 -k1 U: -a0 -mM -k17 -dIL`.

4.2 Unicode vs Wylie Transliteration

The above-mentioned taggers were trained and tested on Tibetan script in Unicode. The Unicode Tibetan script contains a lot of so-called 'stacked' characters that are centred before, above and below one single root letter. A typical example is བསྒྲུབས , which is transliterated in the official Wylie system as *bsgrubs* 'achieved'. In Tibetan Unicode, the order of these stacked characters can differ depending on the exact combinations of consonants and vowels. This varying order often yields unexpected problems when processing Tibetan Unicode text as our NLP algorithms do not recognise variants of the same order in the same word as the same type. This then increases the number of types and thus reduces the overall accuracy. For this reason, we converted the Classical Tibetan Gold Standard from Tibetan Unicode script to Wylie transcription as well. Some examples of Tibetan Unicode with Wylie transliterations are: བཅོམ་ལྡན་འདས་ *bcom-ldan-'das* 'Blessed One', ཤཱཀྱ་སེང་གེ་ *shAkya-seng-ge* 'Buddha', ཕྱག་ *phyag* 'arm, prostration'.

In a 10-fold cross-validation of the Classical Tibetan Gold Standard, this conversion to Wylie yields slightly better results. Global Accuracy was 95.0% for Tibetan Unicode vs. 96.5% for Wylie. We observed a major improvement in Unknown Words in particular from 53.4% in the Tibetan Unicode to 62.2% in the Wylie transliteration. Since the results with the Wylie transliteration are slightly better, especially for unknown, out-of-vocabulary items, converting all Unicode Tibetan to Wylie transliteration would appear to be a logical way forward. However, in practice, Unicode Tibetan script is far more widely used within the Tibetan community. To make the corpus more accessible, but also to get support from members of this community who are willing to correct segmentation and any further type of linguistic annotation, a Unicode Tibetan version is indispensable. It is therefore important to develop segmenters, taggers and parsers that work well for both, or develop tools that can automatically convert the Tibetan text (but not any type of annotation also in roman script) back from its Wylie transliteration to Unicode Tibetan script.

4.3 Memory-Based vs Neural-Network Tagging

Finally, we tested a BiLSTM-CNN-CRF tagger[5] to see if it would yield better results than the memory-based tagger. We chose this neural-network tagger, because it processes both word- and character-level representations automatically, using a combination of a bidirectional Long-Short-Term-Memory (LSTM), a Convolutional Neural-Network (CNN) and a Conditional Random Field (CRF). Although this tagger requires no pre-processing of the data or any further feature engineering, the results are better when the system can use word vectors for the specific language. Since the current number of manually corrected tokens in Old Tibetan is too small to train any neural-network-based tool, we again resorted to using Classical Tibetan instead. For Classical Tibetan, we used FastText[6] to create word embeddings with the aim of improving the results of the tagger with word vectors based on a large amount of Tibetan data digitised by the BDRC[7] and annotated by Meelen and Hill 2017: the Annotated Corpus of Classical Tibetan (ACTib) (version 1, (Meelen et al., 2017)). We then divided the above-mentioned $>318k$ token Classical Tibetan Gold Standard in training, test and developments sets (80/10/10), trained a tagger with these word embeddings and evaluated the results on the held-out test set. With its default settings,[8] this BiLSTM-CNN-CRF tagger yielded a result of 95.8% Global Accuracy (F1 score).[9]

These results are slightly better than those of the memory-based tagger (95% Global Accuracy). They are reasonable, but could be improved in a number of ways. Furthermore, at present they cannot easily be reproduced for our small corpus of the Old Tibetan *Annals* written in a very different style and genre.

[5] See `https://github.com/achernodub/targer` and Chernodub et al. (2019).

[6] `https://fasttext.cc/`

[7] `https://www.tbrc.org/`

[8] Batch size = 10; 100 epochs; dropout ration = 0.5 with the Bi-RNN-CNN-CRF model.

[9] Although it is not common practice (anymore) for POS tagging evaluations, we calculated the F1 instead of normal accuracy to make it directly comparable to the results presented by (Meelen and Hill, 2017). Actual accuracies are slightly higher than the Global Accuracies presented here.

These initial neural-network results thus look promising, but need further extension and refinement. In forthcoming work we address these issues by optimising the parameters, improving the segmentation and, with that, creating better word embeddings (Hill et al., ming).

4.4 Summary of POS Tagging

The below table summarises the results of our tests and evaluations discussed in the previous sections. There are some differences between the small and the larger tag sets and between the Unicode Tibetan script and the Wylie transliteration, with the smaller tag sets and the Wylie transliteration getting better results. The neural network tagger performs best overall with the larger tag set. With the smaller tag set, the Wylie transliteration is best for the smaller tag set.

	Global Accuracy
Clas. Tib. (318k; 15 tags)	96.3%
Old Tib. (3.5k; 15 tags)	92.8%
Old & Clas. (321.5k; 15 tags)	96.1%
Wylie translit. (318k; 15 tags)	96.5%
Unicode Tib. (318k; 79 tags)	95.0%
Wylie translit. (318k; 79 tags)	94.7%
NN-tagger (318k; 79 tags)	95.8%

5 Chunk-Parsing

To facilitate further future research, we also developed a 'hierarchical chunk-parse' of our Old Tibetan corpus. This is a detailed, but rather shallow parse that aims to be as theory-neutral as possible. Constituents are combined into phrases where necessary and uncontroversial, in a hierarchical fashion, e.g. nouns can combine with adjectives and determiners into a Determiner Phrase (DP), which can then combine with a post-positional case marker into a Pre/Postpositional Phrase (PP).

With the small tag set, all case markers are automatically converted into adpositions. This includes the 'Agentive Case' (`case.agn`) that is used to indicate the subject of transitive verbs. If instead we keep this agentive case marker, our small tag set will be extended, but since this marker is highly consistent in spelling, its Precision, Recall and f-score are extremely high (98%, 100% and 99% respectively for n=5627 in the Wylie transliteration evaluation of Classical Tibetan discussed above). The advantage of keeping the agentive case marker tag is that for many transitive sentences at least, we will be able to automatically detect the subject of the clause. Since Old Tibetan was a pro-drop language (i.e. pronouns need not necessarily be overtly expressed, see Tournadre 2010, 101), it is not always possible to detect non-marked subjects of verbs automatically, so a certain amount of manual correction is still always necessary. Similarly, keeping the genitive case markers (ཀྱི་ གྱི་ ཡི་/`case.gen`, see Tournadre and Dorje 2003, 102) has the advantage of getting much better automatically chunk-parsed results for complex nominals.

We used the NLTK chunk-parser[10] to combine tagged tokens into phrases. Semi-hierarchical structures were created by carefully formulating all phrase formation rules in the correct order, e.g. adjectival phrases (`ADJP`) before noun phrases (`NP`) and determiner phrases (`DP`) before pre/postpositional phrases (`PP`). A set of sample rules developed to generate a RegEx grammar for Old Tibetan looks like this:

```
ADJP: {<ADJ><ADJ>?}
NP: {<NOUN|PROPN>}
NUMP: {<NUM><NUM>?}
DP: {<DET>?<NP>?<ADJP|NUMP>?<DET>}
DP: {<NP><ADJP|NUMP><ADJP|NUMP>?}
DP: {<NP|DP><case.gen><NP|DP>}
SbjNP: {<NP|DP><case.agn>}
PP: {<DP|NP><ADP>}
VP: {<VERB|AUX>?<VERB|AUX>}
ADVP: {<ADV><ADV>?}
```

Some sample results are shown in (3) and (4):

(3) (S(SbjNP(NP དཀྱལ་མང་པོའི/PROPN) ས/case.agn)
(PP (NP ཞིང་/NOUN) ཀྱི/ADP)
(NP ཕྱིང་རིལ་/NOUN) (VP བཀྱིས/VERB))
da rgyal mang po rje-s zhing gyi phying ril bgyis
'Dargyal Mangporje carried out a 'felt roll tax'.'

(4) (S(PP(DP(NP ཞང་ཞུང་ཡུལ་/PROPN) གྱི/case.gen)
(NP སྨན་/NOUN)) ཏུ/ADP)
(NP སྤུག་ཙམ་རྩན་རྩ་ཆུང་/PROPN) (VP བཅུག/V)

zhang zhung yul gyi mngan du spug gyim rtsan rma chung bcug

[10] `http://www.nltk.org`

'[He] installed Spug Gyimrtsan Rmachung as the fiscal governor of the land of Zhang-zhung.'

By exploiting the language's standard head-final word order, we can create subordinate clauses for phrases with nominalised verbs ending in subordinate conjunctions. Similarly, we can create relative clauses for nominalised verbs followed by the genitive, which functions as a relative marker linking the following word to the preceding relative clause.[11] The results require only minimal manual correction and are sufficiently theory-neutral to facilitate morpho-syntactic research within a variety of frameworks. The bracket notation is formatted according to the standard `.psd` guidelines and converted to `.psdx` (a TEI XML version of `.psd`) so that they can be queried by CorpusSearch,[12] CorpusStudio[13] or any other plain text or XML-based way of querying syntactic data. These semi-hierarchical structures are not only useful for historical syntacticians interested in comparing basic phrasal structure in different languages, but they are also invaluable for students and scholars of Tibetan to get a good insight into how the grammar of the language has changed over time. Finally, this semi-hierarchical phrasal structure serves as a great starting point for further Old Tibetan NLP challenges, such as creating more meaningful word embeddings and developing tools for keyphrase extraction, document clustering and topic modelling.[14]

6 Post-Processing

Throughout this paper we have shown how automatic NLP tools for Classical Tibetan can be optimised and extended in order to get as much use out of them for Old Tibetan. In this final section we present the results of a thorough error analysis. Suggestions for semi-automatic and rule-based corrections center around Old Tibetan, though some could be extended to the Classical Tibetan data as well.

[11] See Meelen and Roux (fc) for further examples of semi-automatic syntactic annotation.

[12] `corpussearch.sourceforge.net/index.html`

[13] `https://dev.clarin.nl/node/4239`

[14] After finishing all manual corrections at the end of the year, the entire annotated corpus and tools will be made available through Github.

6.1 Correction & Error Analysis

For the segmentation stage clear errors are instances of case markers and converbs that are still attached to the tokens they modify, but these markers should each receive their own tag. Because of their consistent orthography, they can often easily be split from their preceding token to facilitate POS tagging and parsing. In addition, these homophonous forms could be checked after POS tagging: their tag should be a converb following a verb, but a case marker following a noun. Similarly, a simple dictionary look-up script could 'check' whether the forms proposed by the segmenter actually exist. In order to make this latter loop-up task work well, however, we first need to collate and convert Old and/or Classical dictionaries into a reliable and searchable format.

6.1.1 Specific Old Tibetan Errors

We have detected a number of specific Old Tibetan errors as well. In example (5), for instance, we can identify some regular mistakes. Adverbial expressions like དགུན *dgun* 'in winter', དབྱར *dbyard* 'in summer', have been tagged as nouns in many instances, so we can search for these and other recurring adverbial expressions and replace their incorrect nominal tags.

(5) བཙན་པོ་ དབྱར་ སྤེལ་ ན་ བཞུགས་ ཤིང་
NOUN ADV PROPN ADP VERB SCONJ
"In summer, the emperor stayed in Spel."

Furthermore, converbs (functioning like subordinate conjunctions, `SCONJ`) like the ཤིང *shing* 'and, while' have often been tagged as particles instead of subordinate conjunctions, which again, can be automatically replaced.

The large amount of proper nouns in historical texts such as the Old Tibetan *Annals*, however, create a real challenge for our tools. For now, most of the time these tags (and segmentation) had to be corrected manually. For example, the following sentence was originally segmented and tagged as follows:

(6) བྲག་ར་ ན་ བཞུགས །
NOUN ADP ADP VERB PUNCT
Lit: 'cliff into/for in stayed'

The correct analysis here instead should combine the ར -*r*, which was originally tagged as an

adposition (ADP) with the preceding noun དྲག་མ *brag ma* 'cliff', resulting in the proper noun of the place called 'Dagmar':[15]

(7) དྲག་མར་ ན་ བཞུགས །
PROPN ADP VERB PUNCT
"[he] resided in Dagmar"

This correction, as many others occurring with proper nouns, cannot be done automatically since the error patterns are not regular. Sometimes *Dagmar*, a toponym, is tagged correctly as a proper noun, however, *dagma* + *r* 'into a cliff' is also a possible segmentation, in which case the correct POS tags would be NOUN + ADP. Since the Tibetan script does not identify capital letters, it is difficult for any NLP tool to make the right decision in these cases. It would also be difficult to look up ambiguous forms like these in a comprehensive, searchable Old Tibetan proper noun lexicon (which we are currently developing), as the alternative reading is still possible. This issue is exacerbated by the fact that Tibetan proper nouns are almost exclusively also normal nouns, mainly referring to natural phenomena, e.g. *Nyima* 'sun, Nyima'.

6.2 De-Normalisation

Since in the pre-processing stage we converted and normalised our Old Tibetan to 'Classical Tibetan' orthography, in the post-processing stage we need to reverse the Cg3 normalization rules and apply them to the normalised text. This task is straightforward since the Cg3 normalisation grammar has been created with this de-normalisation process in mind. Through selecting and deselecting the OT and σ tags respectively, we converted our Old Tibetan corpus back to its original form after annotation.

7 Conclusion

Developing this Old Tibetan Treebank is a challenging case study of applying NLP tools to extremely low-resourced languages. We overcame many obstacles by first converting/normalising the Old Tibetan to a form of Tibetan that is orthographically much more similar to Classical Tibetan, so that the few extant tools for Classical Tibetan could be tested. We then optimised and extended these tools in various ways and finally developed a chunk-parser to create the first Old Tibetan Treebank as an indispensable tool for philologists, linguist, but also for scholars in Tibetan studies and the Tibetan communities, as it facilitates the development of good Tibetan dictionaries and other Tibetan NLP tools.

[15] The initial consonant cluster *br-* is pronounced as a retroflex /*d*/ in Tibetan, hence the initial *D-* in the place name.

References

Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., and Panchenko, A. (2019). Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association of Computational Linguistics (ACL'2019)*, Florence, Italy.

Daelemans, W., Zavrel, J., van den Bosch, A., and Van der Sloot, K. (2003). Mbt: Memory-based tagger. *Reference Guide: ILK Technical Report-ILK*, pages 03–13.

De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592.

Eckhoff, H., Bech, K., Bouma, G., Eide, K., Haug, D., Haugen, O. E., and Jøhndal, M. (2018). The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52(1):29–65.

El-Haj, M., Kruschwitz, U., and Fox, C. (2015). Creating language resources for under-resourced languages: methodologies, and experiments with Arabic. *Language Resources and Evaluation*, 49(3):549–580.

Galves, C. (2018). The Tycho Brahe Corpus of Historical Portuguese. *Linguistic Variation*, 18(1):49–73.

Garrett, E., Hill, N. W., and Zadoks, A. (2014). A rule-based part-of-speech tagger for Classical Tibetan. *Himalayan Linguistics*, 13(2):9–57.

Green, N., Larasati, S. D., and Žabokrtskỳ, Z. (2012). Indonesian dependency treebank: Annotation and parsing. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 137–145.

Hill, N., Meelen, M., and Roux, E. (forthcoming). Improving the Annotated Corpus of Classical Tibetan (ACTib). *TALLIP, special issue on Asian and Low-Resource Language Information Processing.*

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

Hill, N. W. (2010). An overview of Old Tibetan synchronic phonology. *Transactions of the philological society*, 108(2):110–125.

Hill, N. W. (2012). A note on the history and future of the 'Wylie' system. *Revue d'Etudes Tibétaines*, 23:103–105.

Meelen, M. and Hill, N. (2017). Segmenting and POS tagging Classical Tibetan using a memory-based tagger. *Himalayan Linguistics*, 16(2):64–89.

Meelen, M., Hill, N. W., and Handy, C. (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.

Meelen, M. and Roux, E. (fc). Meta-dating the ACTib.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., and Wallenberg, J. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). In *LREC*, pages 1977–1984.

Taylor, A. and Kroch, A. S. (1994). The Penn-Helsinki Parsed Corpus of Middle English. *University of Pennsylvania*.

Tournadre, N. (2010). The Classical Tibetan cases and their transcategoriality: From sacred grammar to modern linguistics. *Himalayan Linguistics*, 9(2):87–125.

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