Title:
A Rule-based Part-of-speech Tagger for Classical Tibetan

Journal Issue:
Himalayan Linguistics, 13(2)

Author:
Garrett, Edward
Hill, Nathan W., SOAS
Zadoks, Abel

Publication Date:
2014

Permalink:
https://escholarship.org/uc/item/5jv3r0rn

Author Bio:
Lecturer in Tibetan and LinguisticsDepartment of China & Inner Asia and Department of Linguistics

Keywords:
Classical Tibetan, Part-of-speech Tagger

Local Identifier:
himalayanlinguistics_24023

Abstract:
This paper reports on the development of a rule-based part-of-speech tagger for Classical Tibetan. Far from being an obscure tool of minor utility to scholars, the rule-based tagger is a key component of a larger initiative aimed at radically transforming the practice of Tibetan linguistics through the application of corpus and computational methods.

Title:
A Rule-based Part-of-speech Tagger for Classical Tibetan

Journal Issue:
Himalayan Linguistics, 13(2)

Author:
Garrett, Edward
Hill, Nathan W., SOAS
Zadoks, Abel

Publication Date:
2014

Permalink:
http://escholarship.org/uc/item/5jv3r0rn

Author Bio:
Lecturer in Tibetan and LinguisticsDepartment of China & Inner Asia and Department of Linguistics

Keywords:
Classical Tibetan, Part-of-speech Tagger

Local Identifier:
himalayanlinguistics_24023

Abstract:
This paper reports on the development of a rule-based part-of-speech tagger for Classical Tibetan. Far from being an obscure tool of minor utility to scholars, the rule-based tagger is a key component of a larger initiative aimed at radically transforming the practice of Tibetan linguistics through the application of corpus and computational methods.

himalayan linguistics

A free refereed web journal and archive devoted to the study of the languages of the Himalayas

Himalayan Linguistics

A rule-based part-of-speech tagger for Classical Tibetan

Edward Garrett

Nathan W. Hill

Abel Zadoks

SOAS, University of London

Abstract

Although corpus linguistics has been one of the major growth areas in linguistics over the last decades, few have explored Himalayan languages with corpus methods. In Tibetan a large number of raw e-texts are at hand, but but the field lacks tools to access this data efficiently. This paper presents the inner-workings of version 1.0 of a rule-based part-of-speech tagger (stable on 6 January 2014) developed by a research project 'Tibetan in Digital Communication' hosted at SOAS, University of London. For each rule we present the motivation for the rule, a natural language statement of the rule, and a machine readable regular expression version of the rule. At present, the rule-based tagger is being used primarily as a time-saving intervention within our tagging workflow. In the long term, the rule-based tagger will be combined with a statistical tagger to achieve improved results.

Keywords

Tibetan, corpus linguistics, part-of-speech tagging

This is a contribution from *Himalayan Linguistics, Vol. 13(1): 9–57.*
ISSN 1544-7502

Himalayan Linguistics, Vol. 13(1). © Himalayan Linguistics 2014
ISSN 1544-7502

A rule-based part-of-speech tagger for Classical Tibetan

Edward Garrett
Nathan W. Hill
Abel Zadoks
SOAS, University of London

Table of Contents

1 Introduction

This paper reports on the development of a rule-based part-of-speech tagger for Classical Tibetan.*Far from being an obscure tool of minor utility to scholars, the rule-based tagger is a key component of a larger initiative aimed at radically transforming the practice of Tibetan linguistics through the application of corpus and computational methods.

Figure 1: Screen shot of rule suggestions (9 November 2013)

*We gratefully acknowledge the UK's Arts and Humanities Research Council for funding this research as part of the project 'Tibetan in Digital Communication'.

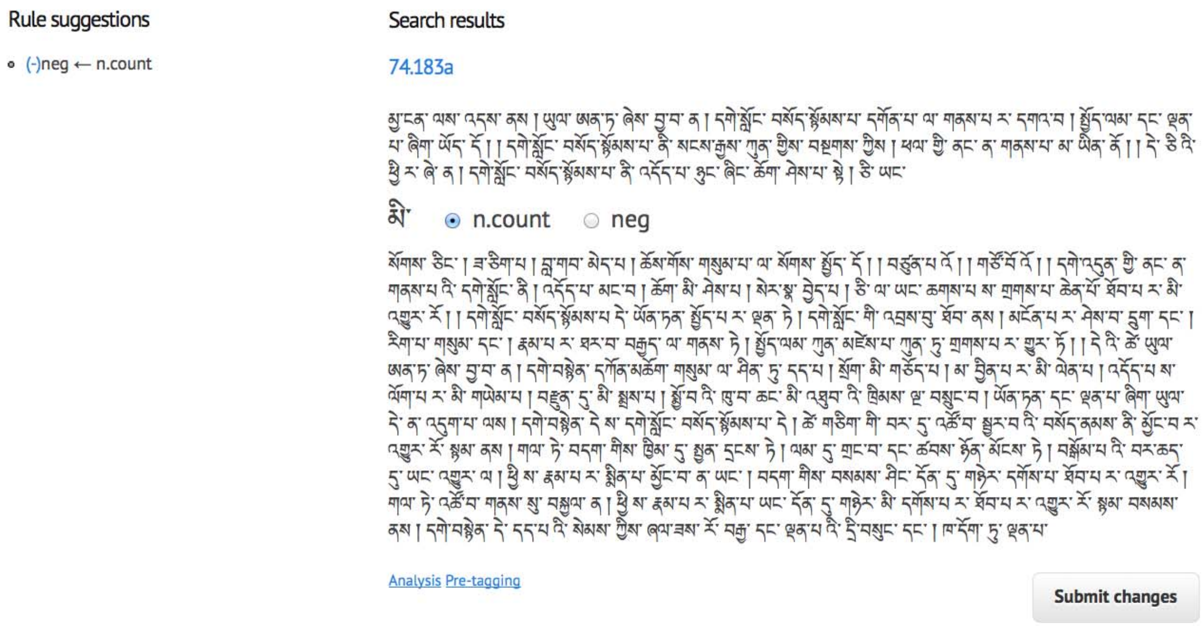

Figure 2: screen shot of the rule suggestion [neg] ← [n.count] (9 November 2013)

Over the years, Tibetology has produced a substantial body of raw electronic data, but the field still lacks in tools to access this data efficiently. The creation of a part-of-speech tagged corpus would open new vistas in Tibetan studies. By allowing for detailed searching to target specific words in particular discourse contexts, it would be the first step in the creation of a historical Tibetan dictionary aimed at meeting the expectations of scientific lexicography, based on corpus linguistics and with examples drawn from attested language use.

The rule-based tagger is currently being used to assist in the compilation of just such a corpus. With help from the tagger, we are creating a 1,000,000 syllable corpus of annotated Tibetan texts, sampled across the whole of Tibetan linguistic history, from the invention of the Tibetan alphabet in 650 CE to the speech of modern Lhasa. This paper focusses on Version 1.0 of the rule-based tagger, for use with Classical Tibetan materials. Subsequent versions of the tagger will be adapted for use with Old and Modern Tibetan.

At present, the rule-based tagger is being used primarily as a time-saving intervention within our tagging workflow. Individual tags must still be hand-checked, but the human annotator's job is considerably simplified through the elimination of impossible tags. With this intervention, the human annotator can focus her attention on the more difficult tagging decisions that the rule-based tagger is unable to disambiguate.

In the long term, the rule-based tagger will be combined with a statistical tagger to achieve improved results. Rule-based approaches parallel the rules of thumb that one might teach a first year Tibetan student (e.g. if *lo* occurs before a *śad* and after a verb stem that ends in *-l*, then it is not the noun 'year'), and are especially effective for rare or systematic phenomena governed by known linguistic generalizations. Statistical approaches, by contrast, parallel an experienced reader's intuitive grasp of a text; the statistical model extracts patterns and regularities from previous exposure to tagged texts, enabling it to choose the most likely interpretation of a new text, without necessarily applying explicitly linguistic knowledge or expertise. As our corpus grows in size, we will incorporate a statistical tagger, which will enable the rule-based tagger to take on a more specialized function.

Our project began by hand tagging an initial 17,522 words of the *Mdzaṅs blun*. We developed the initial part-of-speech tag set during this phase. In the next phase, covering the next 26,937 words of the *Mdzaṅs blun* and the first 32,083 words of the *Mi la ras paḥi rnam thar*, we developed the rule-based tagger through an ad hoc process of trial and error. The rule-based tagger intervenes into the work flow in two moments. First, the output of the rule based tagger on untagged text yields 'pre-tagging' that is referred to a human annotator. The human annotator adjusts the tagging to correct errors. In the course of her work, the human annotator is likely to grow weary of incessantly correcting the same type of mistakes; noting that some of these errors are amendable to rule-based specification, she recommends the addition of further rules to the rule-based tagger. Once complete, the work of the human annotator is fed back into the system. The rule-based tagger, incorporating the newly suggested rules, is now run a second time; cases where the rule-based tagger reaches an unambiguous analysis that differs from the analysis of the human annotator are at this point flagged as 'suggestions'. Each suggestion either reflects an error of the human annotator or an incorrect specification of a rule. The tagging of the corpus or the statement of the rules are modified until there are no more 'suggestions'.

Figure 1 shows how the system displays its overview of the rule suggestions. Figure 2 offers a screen shot of a specific rule suggestion. In this case, seeing the syllable *mi* before a verb, the computer suggests that it is the negation prefix. This time the human annotator is correct and the specification of the rule is not correct. The syllable *mi* is the noun 'man'. Based on the intuition that the verb *sogs* 'etc.' is unlikely to be negated, more recent versions of the tagger preclude this suggestion before this particular verb.

This paper presents the inner-workings of version 1.0 of the rule-based part-of-speech tagger (stable on 6 January 2014). For each rule we present the motivation for the rule, a natural language statement of the rule, and a machine readable regular expression version of the rule.

2 The basic part-of-speech tag set

Before asking what part-of-speech category a particular Tibetan word belongs to, it is necessary to establish the available set of part-of-speech categories. Garrett et al. (forthcoming) describes a part-of-speech tag set for Classical Tibetan developed on the basis of the first 17,522 words of the *Mdzaṅs blun*. An alphabetized list of the current part-of-speech-tag set is presented here with succinct descriptions; Garrett et al. (forthcoming) provides fuller discussion.

[adj] adjectives (e.g. *chen-po* 'big', *bzaṅ-po* 'good', *g.yas-pa* 'right' and *gcig-pa* 'alone' etc.)
[adv.dir] 'directional adverbs' (*phyin-cad* 'after', *sṅon-cad* 'before', *man-cad* 'below', *yan-cad* 'above', *slan-cad* 'after', *phan-tshun* 'mutually')
[adv.intense] 'intensive adverbs' (*rab* [*tu*] 'very', *śin* [*tu*] 'very', *ha-caṅ* 'very')
[adv.proclausal] 'proclausal adverbs' (*de* [*nas*] 'then', *de* [*ste*] 'thereafter', *gal* [*te*] 'if', *ḥo* [*na*] 'in that case', *ḥon* [*te*] 'nevertheless', *yaṅ* [*na*] 'alternatively')
[adv.temp] 'temporal adverbs' (*sṅon* 'previously', *da* 'now', *deṅ* 'these days', *mdaṅ* 'yesterday', *gdod* 'at first', *da-ruṅ* 'still', *phyi-ñin* 'the next day', *phyi-dro* 'in the afternoon', and *saṅ* 'the

next day')
[case.abl] the affix -*las* after a noun phrase
[case.agn] the affixes -*gis, -gyis, -kyis, -s* after a noun phrase
[case.all] the affix -*la* after a noun phrase
[case.ass] the affix -*daṅ* after a noun phrase
[case.comp] the affixes -*bas* and -*pas* after a noun phrase
[case.ela] the affix -*las* after a noun phrase
[case.gen] the affixes -*gi, -gyi, -kyi, -ḥi* and -*yi* after a noun phrase (and in some cases after verbs, e.g. *ḥgyur gyi mi*, *soṅ gi phyir*, etc.)
[case.loc] the affix -*na* after a noun phrase
[case.term] the affixes -*tu, -du, -ru, -su, -r* after a noun phrase
[cl.focus] the focus clitics *ni, kyaṅ, yaṅ, ḥaṅ, caṅ,* and *phyir-yaṅ*
[cl.lta] the clitic *lta* in the combinations *lta ste* and *na lta* (i.e. not *ḥdi ltar*, *lta-bu* etc.)
[cl.quot] the quotative clitics *ces, źes, sñam, źe, ces-pa, ces-pa, źes-pa*
[cl.tsam] the clitics -*tsam*, -*sñed*, -*sñad*
[cv.abl] the affix -*las* after a verb stem
[cv.agn] the affixes -*gis, -gyis, -kyis, -s* after a verb stem
[cv.all] the affix -*la* after a verb stem
[cv.are] the affix -*ta-re* and its allomorphs after a verb stem
[cv.ass] the affix -*daṅ* after a verb stem
[cv.ela] the affix -*las* after a verb stem
[cv.fin] the affixes -*to*, -*no*, -*so*, etc. after a verb stem
[cv.gen] the affixes -*gi, -gyi, -kyi, -ḥi* and -*yi* after a verb stem
[cv.imp] the affixes -*cig*, -*źig*, -*śig* after a verb stem
[cv.impf] the affixes -*ciṅ*, -*źiṅ*, -*śiṅ*
[cv.loc] the affix -*na* after a verb stem
[cv.ques] the affixes -*tam* and its allomorphs.
[cv.sem] the affixes -*te, -de, -ste*
[cv.term] the affixes -*tu, -du, -ru, -su, -r* after a verb stem
[dunno] a word that we have not been able to analyze
[n.count] lexical nouns (e.g. *rgyal-po* 'king', *śiṅ* 'tree', *gaṅ-na-ba* 'whereabouts', *kun-tu-rgyu* 'parivrājaka')
[n.prop] proper nouns (e.g. *Kun-dgaḥ-bo* 'Ānanda', etc.)
[n.rel] relator nouns (e.g. [*deḥi*] *naṅ* [*na*] 'inside of that', [*deḥi*] *druṅ* [*du*] 'before him', [*deḥi*] *ḥog* [*tu*] 'under that', [*deḥi*] *tshe* [*na*] 'at that time', [*ḥdi*] *lta*[*r*] 'like this' etc.)
[n.mass] mass nouns (*nor* 'wealth', *chu* 'water', *zaṅs* 'copper', etc.)
[neg] the two negation prefixes *ma* and *mi*
[num.card] cardinal numbers (e.g. *gcig, gñis, gsum*, etc.)
[num.ord] ordinal numbers (*daṅ-po, gñis-pa, gsum-pa*, etc.)
[p.indef] indefinite pronouns (*la-la* 'some', *so-so* 'each', *gñi-ga* 'both', *gsum-ka* 'the three')
[p.interrog] interrogative pronouns (*su* 'who', *nam* 'when', and *gaṅ* 'where')
[p.pers] personal pronouns (e.g. *ṅa*, *bdag-cag*, *kho-bo*, … *khyod*, *khyed*, etc.)
[d.dem] demonstratives (*ḥdi* 'this', *de* 'that', *phyi*[*r*] 'back, outside')
[d.det] determiners (*gźan* 'other', *ya-re* 'each one (of two)', *ḥbaḥ* 'sole', *śa-stag* 'only', *re* 'respective')

[d.emph] emphatics (*ñid* as in *rgyal-po ñid* 'that very king', *kho-na* 'the very, same', *re-re* 'each')
[d.indef] the indefinite (*cig* etc. as in *pho-ña cig* 'a messenger')
[d.plural] markers of the plural (*rnams, dag, kun, thams-cad, ḥo-cog* [and its variants], *tsho*, *ḥgaḥ* 'some', *sogs* 'etc.')
[v.aux] auxiliary verbs (*nus* 'be able', [*ma*] *thag* 'just, immediately', *srid* 'be possible', *ḥdod* 'want', *ran* 'be time for', *mod* 'indeed')
[v.cop] copula verbs (*yin, lags, mchis*, etc.)
[v.cop.neg] the inherently negative copula verb *min*
[v.neg] the inherently negative verb *med*
[v.pres] present verb stem (*gsod*, *gcod*, [*ma*] *gśegs* [*śig*], etc.)
[v.past] past verb stem (*bsad*, *bcad*, [*ma*] *gśegs* [*so*], *gsol* [*to*], etc.)
[v.fut] future verb stem (*gsad, gcad*, etc.)
[v.imp] imperative verb stem (*sod*, *chod*, *gśegs* [*śig*], etc.)
[n.v.aux] nominalized (*-pa*/*-ba*) equivalent of [v.aux]
[n.v.cop] nominalized (*-pa*/*-ba*) equivalent of [v.cop]
[n.v.cop.neg] nominalized (*-pa*/*-ba*) equivalent of [v.cop.neg]
[n.v.neg] nominalized (*-pa*/*-ba*) equivalent of [v.neg]
[n.v.pres] nominalized (*-pa*/*-ba*) equivalent of [v.pres]
[n.v.past] nominalized (*-pa*/*-ba*) equivalent of [v.past]
[n.v.fut] nominalized (*-pa*/*-ba*) equivalent of [v.fut]
[punc] the punctuation marks །, ༏, ༎, ༈།, and ༎༎

3 The rule-based tagger in action

The rule based tagger functions in two broad phases: it applies as many part-of-speech tags as possible to each word, and then removes deprecated analyses. In the first phase, each word of a text is compared automatically against a digitized version of a verb dictionary (Hill 2010) and the previous body of hand-tagged materials. Any part-of-speech tags found associated with a word in one of these two sources is then supplied to this word. For example, examining the word *chos* the computer finds the analysis [v.imp] in the verb dictionary and the analysis [n.count] in previously hand-tagged materials; it therefore associates both [v.imp] and [n.count] with the instance of *chos* under examination, before moving on to the following word. Eventually all of the words in the text are associated with all of the possible analyses found in both the verb dictionary and in previously tagged text. Figure 3 shows a very short passage as it might appear after this first phase of processing.

After all words in a text are associated with all of their respective part-of-speech analyses the rule-based tagger applies a set of rules one by one to delete out incorrect analyses. In the result many words have only one analysis, presumably correct, but other words have multiple analyses. Figure 4 shows the same short passage as it appears after the second phase of the rule based tagging. The differences between Figure 3 and Figure 4 illustrates the work of the rule-based tagger: after the noun *rgyal-po* the analysis of *de* as the semi-final converb is eliminated.

After all of the rules have been run, the result, 'pre-tagging', is referred to the human user as a vertical list of words and the still remaining possible analyses. The human user deletes out the incorrect analyses before returning the completed text to the computer (cf. Figure 5).

Word	Part-of-speech tag
རྒྱལ་པོ་	n.count
དེ་	d.dem ~ cv.sem
ལ་	case.all ~ n.count
བཙུན་མོ་	n.count
ལྔ་	num.card
བརྒྱ་	num.card
ཡོད་	v.invar
ཀྱང་	cl.focus
།	punc

Figure 3: Look-up of possible analyses

Word	Part-of-speech tag
རྒྱལ་པོ་	n.count
དེ་	d.dem
ལ་	case.all ~ n.count
བཙུན་མོ་	n.count
ལྔ་	num.card
བརྒྱ་	num.card
ཡོད་	v.invar
ཀྱང་	cl.focus
།	punc

Figure 4: Pre-tagging

Word	Part-of-speech tag
རྒྱལ་པོ་	n.count
དེ་	d.dem
ལ་	case.all
བཙུན་མོ་	n.count
ལྔ་	num.card
བརྒྱ་	num.card
ཡོད་	v.invar
ཀྱང་	cl.focus
།	punc

Figure 5: Hand-tagging

4 Additional tags for verb forms with ambiguous tense

Unfortunately, for certain verb forms it is not possible in all cases for the human user to specify an unambiguous tense analysis.[1] In order to present the computer with a one-to-one correspondence of words and part-of-speech tags, it was necessary to create a further eight part-of-speech tags that are used in circumstances when the interpretation of the tenses remains ambiguous.

Word	Part-of-speech tag
རྒྱལ་པོ་	n.count
དེ་	adv.proclausal ~ d.dem ~ cv.sem
མི་	neg ~ n.count
དཀའ་	v.fut ~ v.past ~ v.pres ~ v.imp ~ n.count
ཞིང་	cv.impf ~ n.count
།	punc

Figure 6: Look-up of possible analyses

Word	Part-of-speech tag
རྒྱལ་པོ་	n.count
དེ་	adv.proclausal ~ d.dem
མི་	neg
དཀའ་	v.fut ~ v.pres
ཞིང་	cv.impf ~ n.count
།	punc

Figure 7: Pre-tagging before verb stem ambiguation

1 In this paper the term 'verb stem' is used in opposition to 'verbal noun'. Consequently, 'tense' is used to refer to the distinct four principal parts of verbs used in the indigenous grammatical tradition. This terminology is not intended to imply that the morphosyntactic categories recognized by the indigenous tradition correspond semantically to 'tense' (as opposed to 'aspect' or 'mood') as it is used in linguistic typology.

Word	Part-of-speech tag
རྒྱལ་པོ་	n.count
དེ་	adv.proclausal ~ d.dem
མི་	neg
དགའ་	v.fut.v.pres
ཞིང་	cv.impf ~ n.count
།	punc

Figure 8: Pre-tagging after verb stem ambiguation

Word	Part-of-speech tag
རྒྱལ་པོ་	n.count
དེ་	d.dem
མི་	neg
དགའ་	v.fut.v.pres
ཞིང་	cv.impf
།	punc

Figure 9: Hand tagging

The circumstances giving rise to tense ambiguity are best illustrated with an example. The verb *gśegs* 'go' is invariant across all four tenses. Often syntactic cues disambiguate the correct tense (e.g. *gśegs śig* must be the imperative), but in other contexts disambiguation is not univocal. In the phrase *gśegs nas*, the verb *gśegs* is either a past (cf. *byas nas*) or a present (cf. *byed nas*) but not a future.[2] We introduce the tag [v.past.v.pres] to specify that in this and comparable contexts it is impractical to decide between [v.past] and [v.pres]. Similarly, in the phrase *mi gśegs* the verb *gśegs* is either a present (cf. *mi byed*) or a future (cf. *mi bya*), but cannot be understood as a past. We introduce the tag [v.fut.v.pres] to specify that in this and comparable contexts it is impractical to decide between [v.fut] and [v.pres]. Finally, there are contexts such as *gśegs śiṅ* and *gśegs so,* in which it is only possible to say that *gśegs* is not the imperative (cf. *byed ciṅ*, *bya źiṅ*, *byas śiṅ*; and *byed do*, *byaḥo*, *byas so*). Rather than tagging such contexts with the lengthy [v.fut.v.past.v.pres] we instead employ the tag [v.invar]. One must bear in mind, however, that use of the tag [v.invar] is not a positive claim that a verb is (morphologically or otherwise) invariant, but rather is the negative claim that the stem of this verb in this context cannot be more precisely stated. The four new tags for ambiguous verb stems each has a parallel tag for the corresponding verbal nouns.

[v.fut.v.pres] a verb stem indeterminate between future and present
[v.fut.v.past] a verb stem indeterminate between future and past
[v.past.v.pres] a verb stem indeterminate between past and present
[v.invar] a verb stem indeterminate between future, past, and present
[n.v.fut.n.v.pres] the nominalized equivalent of [v.fut.v.pres]
[n.v.fut.n.v.past] the nominalized equivalent of [v.fut.v.past]
[n.v.past.n.v.pres] the nominalized equivalent of [v.past.v.pres]
[n.v.invar] the nominalized equivalent of [v.invar]

2 All examples of *bya nas* in the Derge Kanjur involve either *bya* 'bird' or *nas* 'barley'.

Figures 6-9 illustrate the work flow after the incorporation of these new tags. Figure 6 shows a short passage with all possible part-of-speech tags associated with every word. Figure 7 shows the results that the rule based tagger achieves in removing incorrect analyses. In addition to excluding the analysis of *mi* as the noun 'person' and the analysis of *de* as the semi-final converb, the system has pared down the possible analyses of *dgaḥ* from five to two. The rule based tagger is unable to decide whether *dgaḥ* is a present or future in this context.

In those cases where the computer cannot decide upon a univocal analysis of a verb's tense, it may be possible for human annotators to determine, on the basis of other factors, whether an indeterminate stem is past, present, or future. However, this is a difficult interpretive task requiring a greater understanding of the text and its context than is to be expected (or desired) during part-of-speech tagging. For example, if the phrase *bdag rab tu dbyuṅ du gsol* 'I request that you give me ordination' occurs in close proximity to *bdag la saṅs-rgyas kyi chos bśad du gsol* 'I request that you explain to me the Buddha's dharma', a reader may reason that because *dbyuṅ* is a morphological future it is plausible to understand *bśad* as future in this context. In order to not prejudice future investigations, in our project the human annotator is not asked to specify verb tense beyond the level achieved by the rule based tagger.

The possibility remains that not all Tibetan verbs have four distinct tenses. Many grammarians believe that a class of verbs never distinguishes present and future, and that this is not a fortuitous ambiguity but rather a meaningful gap (e.g. Beyer 1992: 163-164, Schwieger 2006: 94). If so, the effort to univocally disambiguate tense in every instance is a fool's game.

Returning to the rule-based tagger's treatment of *dgaḥ* in the sequence *mi dgaḥ źiṅ*, the implementation of the ambiguous verb tag [v.fut.v.pres] allows the computer to give this word a single tag, thereby encoding its indeterminacy. Figure 8 shows the same passage after the introduction of ambiguous verb stem tags. The remaining ambiguities, such as whether *źiṅ* is the noun 'field' or the imperfective converb, are referred to the human user for adjudication. Figure 9 presents the final outcome of the hand-tagging of this passage, exactly as annotated text is stored in the online system.

5 Overview of the rule-based tagger's inner-workings

The rule-based tagger operates as an ordered sequence of rules applied to an input text. Input texts must follow a specific structure in order for the rules to apply correctly. The first requirement is that words should be separated from each other by whitespace. (Figure 10 replaces the space with a new line for a cleaner presentation.) Each word itself has two parts, separated by the delimiter |. On the left of the delimiter is the word form itself, and on the right are all possible part-of-speech tags for the word in alphabetical order. Individual tags are contained within brackets, e.g. [n.count], which improves readability and makes the rules easier to formulate.

སྐོང་བ་|[n.count][n.v.fut][n.v.past][n.v.pres]

དམིགས་བུ་|[n.count]

ལ་|[case.all][cv.all][dunno][n.count]

རྗེན་པ|[n.v.pres]

འམ|[cv.ques]

།|[punc]

Figure 10: Input text

The rules use regular expressions to scan the input text, substituting each occurrence of a specific pattern with a specific replacement string. The rules exploit 'capturing groups'[3] to copy parts of the input into the output. Usually, the replacement string only slightly modifies the input match: in most cases, the effect of a rule is to remove one or more possible tags from a word. Since the rule-based tagger is integrated into a workflow based on the Java programming language, the rules are written using Java's regular expressions syntax.[4]

Because the output of some rules feeds into other rules, it is important not only to specify rules correctly but also to put the rules in an optimal order. The first set of rules are of a preparatory nature; they aim to avoid errors that might otherwise occur (§6). Rules 1 to 4 decompose mixed verbs tags into their constituent parts, so that the computer does not proliferate beyond four the number of possible Tibetan verb stems. Rules 5 and 6 avoid possible mistakes in the training data from proliferating during pre-tagging, by constraining verb stems to monosyllables and verbal nouns to disyllables. Rule 7 removes the 'dunno' tag; presenting the human user with 'dunno' as a possible analysis would be pointless since it is equivalent to providing no analysis at all.

Once the preliminary rules have run their course, the subsequent rules apply to strip off incorrect tags. Rules that strip off incorrect analyses isolate three broad classes of phenomena. The first set of rules isolates words that are unambiguous in contexts which are easy to find and would cause problems for subsequent rules if left unspecified; once isolated these words allow subsequent rules to make use of a larger number of unambiguous words (§7). The second set of rules distinguish words into major part-of-speech categories (§8). The third set of rules reconsider verb stems and verbal nouns that according to the lexical resources have more than one tense interpretation, and excludes as many of these interpretations as possible, effectively assigning tenses to portmanteau morphemes (§10).

The first set of rules that strip off incorrect analyses (§7) establishes an infrastructure of secure analyses. These rules themselves fall into three categories. The first group disambiguates a grab-bag

3 A capturing group is a sub-expression in parentheses, which is accessed using $ followed by a numeral. The numeral corresponds to the number of groups in the larger expression reading left to right.

4 See http://docs.oracle.com/javase/6/docs/api/java/util/regex/Pattern.html

of frequent words in certain relatively common fixed combinations (§7.1). Isolating and resolving idiosyncrasies early on protects them from subsequent rule application. Rules 8 to 13 attempt to isolate such idiosyncrasies. For example, the syllable *rtsa* has interpretations as a noun 'root' and a morpheme that is used in the formation of numerals. If *rtsa* occurs between two numerals it is very unlikely to be the noun 'root'. To add a rule that removes the [n.count] tag in these contexts spares the human annotator from having to delete each case manually (cf. rule 13). The second group of rules isolates proclausal adverbs (§7.2, rules 14-17). Each proclausal adverb has another possible reading, e.g. *de* [adv.proclausal] *nas* [case.ela] 'then', versus *de* [d.dem] *nas* [case.ela] 'from there, from him'. Using the fact that proclausal adverbs normally begin a sentence, rules 16 and 17 remove other analyzes in this context. The third group of rules (§7.3) identify sandhi determined converbs (rules 20-23), specifying for example that if *lo* is not preceded by a word that end in *-l* then it cannot be the final converb.

Once those words that are easy to disambiguate in certain contexts have been disambiguated, there is an infrastructure of unambiguous tags to permit the classification of major word classes (§8); this is done in four stages: distinguishing verbs from nouns (24-28), distinguishing negation from nouns (29-35), disambiguating case markers and converbs from other things (39-43), and distinguishing case and converbs from each other (44-47).

Only after verbs have been identified as verbs is it possible to address the question of what tense a particular verb form exhibits. The majority of rules in the tagger work to select the correct verb tense in different contexts (§10). This selection is achieved in three phases: disambiguation (53-64), consolidation of systematic ambiguities (66-69), and re-ambiguation of stems that belong to distinct verbs (70-80). The first of these phases, contextual disambiguation of the four verb stems, itself proceeds in three steps: using the following converbs (53-56), using negation (57-58), and using the presence or absence of the *da-drag* (59-64). In the second phase, having done all that we know how to do in order to disambiguate verb stems, the remaining ambiguities are rewritten with tags that consolidate the ambiguity so that they can be saved in the system (66-69), e.g. *mi* [neg] *gśegs* [v.fut] ~ [v.pres] is replaced with *mi* [neg] *gśegs* [v.fut.v.pres]. The consolidation of ambiguities has a downside; when a single form might belong to two distinct verbs, these consolidated tags efface distinctions which should be preserved. The next phase, that of re-ambiguation (rules 70-83) restores these distinctions. For example, the second phase will change *źu* [v.fut] ~ [v.past] ~ [v.pres] into *źu* [v.invar], but *źu* [v.fut] [v.pres] belong to the verb 'request' whereas *źu* [v.past] belongs to the verb 'melt'; because the human user will want to be presented with *źu* [v.past] ~ [v.past.v.pres] a specific rule must be created to achieve this. Each orthographic form that could belong to separate verbs must be individually specified. We only re-ambiguate the orthographic forms that the first 40,366 words of the *Mdzaṅs blun* present for consideration.

The final group of rules (§11) includes two unrelated rules (84 'Precluding *la* as a noun between two imperatives' and 85 'Finding numbers'), which it is not convenient to run earlier.

6 Avoiding errors

Before the intellectual work of disambiguating different possible part-of-speech tags in different contexts begins, it is convenient to preclude several types of errors. Decomposing mixed stem verb stem and verbal noun tags (such as [v.fut.v.pres], [n.v.invar], etc.) avoids the system treating these as new types of verb tags (§6.2). Constraining verb stems to monosyllables and verbal nouns to disyllables prevents mistakes in the training data from proliferating during pre-tagging (§6.2). Deleting the 'dunno' tag prevents the system from treating a failure to explain something as a possible explanation of it (§6.3).

6.1 Avoiding errors by decomposing mixed [v] and [n.v] tags

Although mixed tags such as [v.past.v.pres] and [v.fut.v.pres] are intended to express an ambiguity, i.e. lack of analysis, there is no way for the computer *a priori* to treat them as structurally different from other tags. The default approach of the computer is to treat [v.past.v.pres] as a new type of verb stem, different from both [v.past] and from [v.pres]. The presence of phrases like "*gśegs* [v.past.v.pres] *nas* [cv.ela]" in the training corpus will lead to *gśegs* [v.past.v.pres] entering the lexicon. As a result, the rule based tagger would naturally ask itself meaningless questions like 'is *gśegs* in this context to be tagged [v.past], [v.pres], or [v.past.v.pres]?'. Decomposing mixed tags before running any other rules of the rule based tagger avoids this risk.

(1). Decomposing the tags [v.invar] and [n.v.invar]

BACKGROUND: The tag [v.invar] is used for verb stems that cannot be disambiguated among future, past, and present; for example, in the phrase *gśegs so* the verb *gśegs* could be any tense (cf. present *byed do*, past *byas so*, and future *byaḥo*). A rule replaces each [v.invar] with "[v.fut] ~ [v.past] ~ [v.pres]". An exactly parallel argument applies for [n.v.invar].

RULE: Replace [v.invar] and [n.v.invar] with "[v.fut] ~ [v.past] ~ [v.pres]" and "[n.v.fut] ~ [n.v.past] ~ [n.v.pres] " respectively.

```
PATTERN:
(\S+\|(?:\[(?!(?:n\.)?v\.)[^\]]*\])*(?:\[(?:n\.)?v\.aux\])?(?:\[(?:n\.)?v\.cop\])?)(?:
\[(?:n\.)?v\.fut\])?(?:\[(?:n\.)?v\.fut\.(?:n\.)?v\.past\])?(?:\[(?:n\.)?v\.fut\.(?:n\
.)?v\.pres\])?(\[(?:n\.)?v\.imp\])?\[(n?\.?v\.)invar\](?:\[(?:n\.)?v\.past\])?(?:\[(?:
n\.)?v\.past\.(?:n\.)?v\.pres\])?(?:\[(?:n\.)?v\.pres\])?(\S*)

REPLACE: $1[$3fut]$2[$3past][$3pres]$4
```

(2). Decomposing the tags [v.fut.v.past] and [n.v.fut.n.v.past]

BACKGROUND: The tag [v.fut.v.past] is used for verb stems that cannot be disambiguated between future and past; for example at the end of a sentence (i.e. before a *śad*) the verb form *bsgyur* is either a future (cf. *bya*|) or a past (cf. *byas*|). A rule replaces each [v.fut.v.past] with "[v.fut] ~ [v.past]". An exactly parallel argument applies for [n.v.fut.v.n.past].

RULE: Replace [v.fut.v.past] and [n.v.fut.n.v.past] with "[v.fut] ~ [v.past]" and "[n.v.fut] ~ [n.v.past]" respectively.

```
PATTERN:
(\S+\|(?:\[(?!(?:n\.)?v\.)[^\]]*\])*(?:\[(?:n\.)?v\.aux\])?(?:\[(?:n\.)?v\.cop\])?)(?:
\[(?:n\.)?v\.fut\])?\[(n?\.?v\.)fut\.(?:n\.)?v\.past\]((?:\[(?:n\.)?v\.fut\.(?:n\.)?v\
.pres\])?(?:\[(?:n\.)?v\.imp\])?)(?:\[(?:n\.)?v\.past\])?((?:\[(?:n\.)?v\.past\.(?:n\.
)?v\.pres\])?(?:\[(?:n\.)?v\.pres\])?\S*)
```

```
REPLACE: $1[$2fut]$3[$2past]$4
```

(3). Decomposing the tags [v.fut.v.pres] and [n.v.fut.n.v.pres]

BACKGROUND: The tag [n.v.fut.n.v.pres] is used for verb stems that cannot be disambiguated between future and present; for example, in the phrase *mi* [neg] *gśegs* the verb *gśegs* could either present (cf. *mi byed*) or future (cf. *mi bya*).[5] A rule replaces each [v.fut.v.pres] with "[v.fut] ~ [v.pres]". An exactly parallel argument applies for [n.v.fut.n.v.pres].

RULE: Replace [v.fut.v.pres] and [n.v.fut.n.v.pres] with "[v.fut] ~ [v.pres]" and "[n.v.fut] ~ [n.v.pres]" respectively.

```
PATTERN:
(\S+\|(?:\[(?!(?:n\.)?v\.)[^\]]*\])*(?:\[(?:n\.)?v\.aux\])?(?:\[(?:n\.)?v\.cop\])?)(?:
\[(?:n\.)?v\.fut\])?\[(n?\.?v\.)fut\.(?:n\.)?v\.pres\]((?:\[(?:n\.)?v\.imp\])?(?:\[(?:
n\.)?v\.past\])?(?:\[(?:n\.)?v\.past\.(?:n\.)?v\.pres\])?)(?:\[(?:n\.)?v\.pres\])?(\S*
)

REPLACE: $1[$2fut]$3[$2pres]$4
```

(4). Decomposing the tags [v.past.v.pres] and [n.v.past.n.v.pres]

BACKGROUND: The tag [v.past.v.pres] is used for verb stems that cannot be disambiguated between past and present; for example, in the phrase *gśegs nas* [cv.ela], the verb *gśegs* is either a past (cf. *byas nas*) or a present (cf. *byed nas*). A rule replaces each [v.past.v.pres] with "[v.past] ~ [v.pres]". An exactly parallel argument applies for [n.v.past.n.v.pres].

RULE: Replace [v.past.v.pres] and [n.v.past.n.v.pres] with "[v.past] ~ [v.pres] " and "[n.v.past] ~ [n.v.pres]" respectively.

```
PATTERN:
(\S+\|(?:\[(?!(?:n\.)?v\.)[^\]]*\])*(?:\[(?:n\.)?v\.aux\])?(?:\[(?:n\.)?v\.cop\])?(?:\
[(?:n\.)?v\.fut\])?(?:\[(?:n\.)?v\.imp\])?)(?:\[(?:n\.)?v\.past\])?\[(n?\.?v\.)past\.(
?:n\.)?v\.pres\](?:\[(?:n\.)?v\.pres\])?(\S*)

REPLACE: $1[$2past][$2pres]$3
```

6.2 Avoiding errors by constraining word structure

Constraining verb stems to monosyllables and verbal nouns to disyllables prevents mistakes in the training data from proliferating during pre-tagging.

(5). Limiting verb stems to single syllable

BACKGROUND: In our understanding of Tibetan morphosyntax all verb stems are monosyllabic. Thus, if the rule based tagger suggests tagging a two or more syllable word as a verb stem, this must have been introduced via a mistake in the training data.

RULE: If a word has more than one syllable then delete all [v.xxx] tags from it.

5 Both *ma gśegs* and *ma byas* are unambiguous pasts

```
PATTERN: (\S+་\S+\|\S*)(?:\[v\.[^\]]*\])+(\S*)

REPLACE: $1$2
```

(6). Limiting verbal nouns to disyllables
BACKGROUND: If verb stems consist always of single syllable, then it follows automatically that verbal nouns must consist of disyllables, the first syllable of which is a verb stem, and the second syllable of which is the nominalization suffix that takes the forms *-pa* and *-ba*. Later documents such as the *Mi la ras paḥi rnam thar* have other verbal noun suffixes such as *-mkhan*, *-sa*, and *-tshul*.

RULE: If a word has more than two syllables remove the analysis [n.v.xxx].

```
PATTERN:
((?:^|\s)(?![^་]+་(?:པ|བ|རྒྱུ|ཐབས|ལུགས|གྲབས|ཚུལ|ཚད|མཁན|ས)་?\|)\S+\|\S*)(?:\[n\.v\.[^\]]*\])+(\S*)

REPLACE: $1$2
```

6.3 Avoiding errors by removing the 'dunno' tag

(7). Removing the 'dunno' tag
BACKGROUND: We use the tag [dunno] for words that we are not yet prepared to assign with a part-of-speech tag. For the rule-baed tagger to suggest [dunno] as an analysis would be equivalent to offering no analysis at all; the presence of [dunno] associated with some words would interfere with the correct performance of rules that make uses of unambiguous contexts. Consequently, we remove [dunno] wherever another analysis is available.

RULE: Remove [dunno] if there are other tags.

```
PATTERN: (\S+\|)(?:(\S+)\[dunno\]|\[dunno\](\S+))(\S*)

REPLACE: $1$2$3$4
```

7 An infrastructure of unambiguous tags

Before systematic disambigution of major form classes (such as nouns versus verbs) can take place, it is necessary to pin down a few words as unambiguous. Some words can be disambiguated with less context than others. By treating those words that require less context first, these words can feed into the rules that analyse those words that require more context.

7.1 Idiosyncratic rules that are used to disambiguate frequent words in certain relatively common fixed combinations

The rules in this section aim to isolate the correct analysis of words that do not constitute a meaningful or coherent set. Instead, these words happen for one reason or another to be amenable to easy disambiguation.

(8). Disambiguating *graṅs* [n.count] and *graṅs* [v.pres]

BACKGROUND: The syllable *graṅs* can be both a noun [n.count] 'number' or an alternate present of the verb *bgraṅ* 'count'. The ambiguity continues with *mi graṅs*, which could either be 'a number (of) people' or 'not counting'. However, if *graṅs* is followed by *med-pa* then it forms a small clause meaning 'numberless' and *mi graṅs med-pa* means 'numberless people'. Thus, it is possible to write a rule that disambiguates *graṅs* in this context.

RULE: Assign *graṅs* the interpretation [n.count] when it occurs directly before *med-pa*

```
PATTERN: ((?:^|\s)གྲངས་)\|\S*\[n\.count\]\S*(\s+མེད་པ་?\|)
```

```
REPLACE: $1|[n.count]$2
```

(9). Disambiguating *skad* [n.rel] and *skad* [n.count]

BACKGROUND: The sequence *skad* has the possible tags [n.count] and [n.rel]. In the very frequent expression *ḥdi skad ces*, it should always be tagged as [n.rel].

RULE: In the phrase *ḥdi skad ces* tag *skad* as [n.rel].

```
PATTERN: (འདི་\|\[d\.dem\]\s+སྐད་)\|\S+\s+((?:ཅེས་?)\|\[cl\.quot\])
```

```
REPLACE: $1|[n.rel] $2
```

(10). Disambiguating *skad* [n.rel] and *skad* [n.count] and *de* [d.dem] from *de* [cv.sem]

BACKGROUND: The sequence *de* has the possible tags [d.dem] and [cv.sem]. The sequence *skad* has the possible tags [n.count] and [n.rel]. In the very frequent expression *de skad smras* the sequence *de* is always [d.dem], the sequence *skad* is always [n.rel],and the sequence *smras* is always [v.past].

RULE: Specify that the sequence *de skad smras* is *de* [d.dem] *skad* [n.rel] *smras* [v.past].

```
PATTERN: དེ་\|\S+\s+སྐད་\|\S+\s+(སྨྲས་?)\|\S+
```

```
REPLACE: དེ་|[d.dem] སྐད་|[n.rel] $1|[v.past]
```

(11). Isolating *lta* [n.rel]

BACKGROUND: The form *lta* can have several possible tags, including [n.rel] and [v.pres]. When *lta* appears in *de lta r*, *ji lta r*, or *ḥdi lta r* then it is unambiguously [n.rel]. In addition the <r(a)> ར་, which has the possible tags [n.count], [case.term], and [cv.term] can be specified as [case.term].

RULE: Assign *lta* the tag [n.rel] and assign <r(a)> ར་ the tag [case.term] in the contexts *de lta r*, *ji lta r*, and *ḥdi lta r*.

```
PATTERN: ((?:^|\s)(?:དེ་|ཇི་|འདི་)\|\S+)\s+ལྟ\|\S+\s+(ར་?)\|\S+
```

```
REPLACE: $1 ལྟ|[n.rel] $2|[case.term]
```

(12). Isolating *chos* [n.count]
BACKGROUND: The sequence *chos* has among its possible tags [n.count] and [v.imp]. In the frequent sequence *saṅs-rgyas kyi chos* it is an unambiguously [n.count].[6]

RULE: Assign *chos* the tag [n.count] when it occurs after *saṅs-rgyas kyi*.

```
PATTERN: ((?:^|\s)སངས་རྒྱས་\|\S+\s+ཀྱི་\|\S+)\s+(ཆོས་?)\|\S+
```

```
REPLACE: $1 $2|[n.count]
```

(13). Isolating morphemes used in the formation of numerals
BACKGROUND: Some syllables occur both as nouns and in the formation of numerals (e.g. *rtsa* 'vein' and *so* 'tooth' versus *sum-cu rtsa gsum* 'thirty three' and *sum-cu so lṅa* 'thirty five'). Between two numbers such syllables require the interpretation [num.card]; in this context other interpretations can be excluded.

RULE: If any word has two possible part-of-speech tags, one of which is [num.card], and this word occurs between two words with the part-of-speech tag [num.card], then assign this word the tag [num.card].

```
PATTERN: (\S+\|\[num\.card\])\s+(\S+)\|\S*\[num\.card\]\S*\s+(\S+\|\[num\.card\])
```

```
REPLACE: $1 $2|[num.card] $3
```

7.2 Finding the proclausal adverbs

The rules in this section aim to isolate the proclausal adverbs. These words are fairly easy to isolate because of their restricted syntactic distribution. In addition, because the syllable *de* has two very frequent analyses (viz. [d.dem] and [cv.sem]), precluding the analysis of this words as [adv.proclausal] in as many contexts as possible will serve to increase the accuracy of the rule-based tagger overall.

(14). Disambiguating *de* [d.dem] from *de* [adv.proclausal]
BACKGROUND: The demonstrative *de* frequently appears at the end of noun phrases, but before case morphology; this is a context in which *de* is not interpretable as a proclausal adverb. Thus, isolating *de* at the end of noun phrases allows the analysis as a proclausal adverb to be excluded. We exclude *nas* [case.ela] from the search, because *de* [adv.proclausal] occurs frequently before *nas* [case.ela].

RULE: If *de* occurs after [adj], [d.xxx], [n.xxx], [num.xxx], or [p.xxx] and before [case.xxx] other than [case.ela], then remove from *de* the analysis [adv.proclausal].

6 An anonymous reviewer recommends changing this rule to the more general specification that *chos* is a noun if it follows an unambiguous noun followed by any form of the genitive. We shall incorporate this suggestion into a future version of the tagger.

```
PATTERN:
(\S+\|(?:\[(?:adj|(?:d|n|num|p)\.[^\.\]]*)\])+\s+དེ་?\|\S*)\[adv\.proclausal\](\S*\s+\S+
\|\S*\[case\.(?!ela)[^\]]*\]\S*)
```

```
REPLACE: $1$2
```

(15). Disambiguating *de* [cv.sem] from *de* [adv.proclausal]

BACKGROUND: The semi-final converb occurs at the end of clauses, i.e. often after a verb stem and before a *śad*; this is a context in which *de* is not interpretable as a proclausal adverb. Thus, isolating *de* after verb stems but before *śad* allows the analysis as a proclausal adverb to be excluded.

RULE: If *de* occurs after [v.xxx] and before ། remove from *de* the analysis [adv.proclausal].

```
PATTERN: (\S+\|\S*\[v\.[^\]]*\]\S*\s+དེ་?\|\S*)\[adv\.proclausal\](\S*\s+།\|\S*)
```

```
REPLACE: $1$2
```

(16). Isolating *ḥo na* [adv.proclausal]

BACKGROUND: Because proclausal adverbs are normally found at the beginning of sentences, and sentences normally end with a *śad* (or a *-g* not followed by a *tsheg*) most proclausal adverbs will occur after a *śad* (or a *-g* not followed by a *tsheg*). In Classical Tibetan *ḥo na* is essentially always a proclausal adverb [adv.proclausal]. Theoretically however, the syllable *ḥo* could be a demonstrative pronoun [d.dem]. Nonetheless, after a *śad* the interpretation of *ḥo* as a demonstrative will be exceedingly rare. Consequently it is prudent to interpret all instances of *ḥo na* which occur after ། to be proclausal adverbs.

RULE: In the sequence ། *ḥo na* tag *ḥo* as [adv.proclausal].

```
PATTERN: (།\|\S+\s+འོ་)\|\S+\s+(ན་)\|\S+
```

```
REPLACE: $1|[adv.proclausal] $2|[case.loc]
```

(17). Isolating *gal* [adv.proclausal]

BACKGROUND: The syllable *gal* should always be tagged as [adv.proclausal] when it occurs before *te*. Some readers might wonder whether *gal te* is not best treated as a single word. However, the *te* here is the usual [cv.sem], so it is best to treat *gal* as an independent word.[7]
RULE: Tag *gal te* as *gal* [adv.proclausal] *te* [cv.sem].

```
PATTERN: (\S+\|\[punc\]\s+གལ་)\|\S+\s+(ཏེ་?\|\[cv\.sem\])
```

```
REPLACE: $1|[adv.proclausal] $2
```

---

7 The other proclausal adverbs (e.g. *ḥo na* or *de nas*) refer semantically to the preceding clause. In contrast *gal te* anticipates a following *na* [cv.loc]. This semantic difference does not however warrant a new part-of-speech tag. There are computational disadvantages to adding new part-of-speech tags, and there are no analytic advantages offered by part-of-speech categories with only one member, since the lexical content of the word itself serves as an adequate means to locate the word and study its behavior.

(18). Isolating *la* [adv.proclausal] and *la* [n.count]

BACKGROUND: The syllable *la* has many interpretations: the allative case, the allative converb, the stem of the proclausal adverb *lar* 'moreover', and the noun 'mountain pass'. At the beginning of a sentence (i.e. after a *śad* or *-g* without a *tsheg*) proclausal adverbs are frequent, and a noun 'mountain pass' is possible. In contrast, since they have to follow something, case markers and converbs are precluded in this position.

RULE: If a word *la* appears after ། (or *-g* without a *tsheg*), then delete [case.all] and [cv.all] from this *la*.

```
PATTERN: (\S*(?:ག|།)\|\S+\s+ལ་\|\S*)\[case\.all\](\S*)\[cv\.all\](\S*)
```

```
REPLACE: $1$2$3
```

(19). Precluding *la* [adv.proclausal] at the end of clauses

BACKGROUND: The syllable *la* has many interpretations: the allative case, the allative converb, the stem of the proclausal adverb *lar* 'moreover', and the noun 'mountain pass'. At the end of a clause (i.e. after a verb or verbal noun but before a *śad* or *-g* without a *tsheg*) the pro-clausal adverb' can be precluded.

RULE: If a word *la* appears after [v.xxx] or [n.v.xxx] and before ། (or *-g* without a *tsheg*), then delete [adv.proclausal] from this *la*.

```
PATTERN: (\S+\|(?:\[(?:n\.)?v\.[^\]]*\])+\s+ལ་?\|\S*)\[adv\.proclausal\](\S*\s+།\|\S+)
```

```
REPLACE: $1$2
```

7.3 Identifying sandhi determined converbs

In some cases a converb happens to coincide with a noun orthographically. The following rules seek to correctly isolate the few cases in which the syllable in question is the noun and not the converb.

(20). Isolating the final converb

The final converb is formed by repeating the last phoneme of the preceding word and adding *-o*. Consequently, the initial consonant of the final converb generally coincides with the final consonant of the preceding word. This sandhi context allows for straightforward identification of the final converbs. However, one must keep in mind that not all morphemes of the correct structure that occur in the correct sandhi context will be [cv.fin]. For example, one might imagine a sentence *khos so bcag* 'he broke teeth', in which a search for the final converb using the sandhi context *-s so* would yield a false positive. The interpretation [cv.fin] is particularly plausible at the end of a sentence, i.e. before *śad* (or equivalently the syllable *-go* not followed by a *tsheg*), or the syllables *źes*, *sñam*, or *zer*.

20a. Finding the final converb using sandhi and sentence breaks

BACKGROUND: The coincidence of correct sandhi phenomena and the end of a sentence essentially guarantees the successful identification of the final converb.

RULE: If C*o* (e.g. *lo*) is preceded by a word that ends with -C (e.g. -*l*) and occurs before a །, *źes*, *sñam* or *zer*, then assign tag [cv.fin] to C*o*.

```
PATTERN: (\S+(\S)་\|\S*\s+\2\u0F7C་?)\|\S+\s+((?:།|ཞེས|སྙམ|ཟེར)་?\|\S*)
```

```
REPLACE: $1|[cv.fin] $3
```

20b. Finding the final converb -go before sentence breaks

BACKGROUND: The allomorph -*go* of the final converb is not used before a *śad*, but instead is used equivalently not followed by a *tsheg*. Consequently, this allomorph requires its own rule.

RULE: If g*o* is preceded by a word that ends with -*g* and is not followed by a *tsheg* then assign the tag [cv.fin] to *go*.

```
PATTERN: (\S+ག་\|\S*\s+གོ)\|\S+
```

```
REPLACE: $1|[cv.fin]
```

20c. Finding the final converb - *ḥo* before sentence breaks

BACKGROUND: The allomorph -*ḥo* of the final converb occurs after verbs that end in open syllables. Rule 20a, because it relies on the reduplication found in all other allomorphs of this morphemes, will not locate the allomorph -*ḥo*. This allomorph requires its own rule. Because it is difficult to specify 'ends with a vowel' when treating Unicode Tibetan, we assume that all occurrences of -*ḥo* before a *śad*, *źes*, *sñam* or *zer* are the final converb.

RULE: If *ḥo* occurs before a །, *źes*, *sñam* or *zer*, then assign the tag [cv.fin] to *ḥo*.

```
PATTERN: ((?:^|\s)འོ་?)\|\S+\s+((?:།|ཞེས|སྙམ|ཟེར)་?\|\S*)
```

```
REPLACE: $1|[cv.fin] $2
```

20d. Finding words that are homophonous with forms of the final converb

BACKGROUND: Candidates for analysis as final convebs that fail to occur in the correct sandhi context can be confidently precluded from this analysis.

RULE: Remove the tag [cv.fin] from all instances of C*o* (e.g. *lo*, but excluding *ḥo*) for which the preceding word does not end with -C (e.g. -*l*).

```
PATTERN: (\S*(\S)་(?:\|\S+)?\s+(?!(?:\2|འ))\S\u0F7C་?\|)(?:\[cv\.fin\](\S+)|(\S+)\[cv\.fin\](\S*))
```

```
REPLACE: $1$3$4$5
```

(21). Isolating the question converb

BACKGROUND: The same sandhi contexts that applied to the final converb also occur for the question converbs. Consequently, a very similar pair of rules can isolate both secure examples of the question converbs and secure examples of words that happen to coincide with the question converb (e.g. *nam* 'when').

21a. Finding the question converb using sandhi and sentence breaks
RULE: If a word of the shape C*am* is preceded by a word that ends with 'C' and occurs before a །, or *źes* or *sñam* or *zer*, then assign tag [cv.ques] to the word C*am*.

```
PATTERN: (\S+(\S)་\|\S*\s+\2མ་?)\|\S+\s+((?:།|ཞེས|སྙམ|ཟེར)་?\|\S*)
```

```
REPLACE: $1|[cv.ques] $3
```

21b. Finding words that are homophonous with forms of the question converb
RULE: Remove tag [cv.ques] from C*am* if preceding word does not end with 'C'.

```
PATTERN: (\S*(\S)་\|\S+\s+(?!\2)\Sམ་?\|)(?:\[cv\.ques\](\S+)|(\S+)\[cv\.ques\](\S*))
```

```
REPLACE: $1$3$4$5
```

(22). Distinguishing *de* [cv.sem] from *de* [d.dem]
BACKGROUND: The syllable *de* can be a demonstrative, a proclausal adverb, or a form of the semi-final converb. As a semifinal converb *de* is one of three phonologically determined allomorphs along with *te* and *ste.* The allomorph *de* of the semifinal converb occurs only after words that end with -*d*. Consequently, any instance of *de* that occurs in other sandhi contexts must be the demonstrative or the proclausal adverb and not the semifinal converb.

RULE: If *de* does not occur immediately after a word that ends in -*d* remove from it the interpretation [cv.sem].

```
PATTERN: (\S+(?<!\Sད་)\|\S+\s+དེ་?\|\S*)\[cv\.sem\](\S*)
```

```
REPLACE: $1$2
```

(23). Isolating the semi-final converb before *śad*
BACKGROUND: The previous rule (22) prohibited the interpretation of *de* as a semi-final converb in incorrect sandhi contexts, but it is difficult to find contexts in which to prohibit the interpretation of *de* as a demonstrative. Although the semi-final converb is frequent after verbs, any *de* after a verb might belong to the following clause as a demonstrative. However, if *de* stands immediately before a *śad*, then its interpretation as belonging to the following clause is unlikely. Consequently, a search for *de* after a verb stem and before *śad*, should yield the semi-final converb.

RULE: If a word with the hypothesized tags [d.dem] and [cv.sem] occurs after a word with an unambiguous verb tag [v.xxx], and before །, then delete the tag [d.dem] from this word.

```
PATTERN: (\S+\|(?:\[v\.[^\]]*\])+\s+\S+\|\S*\[cv\.sem\]\S*)\[d\.dem\](\S*\s+།\|\S+)
```

```
REPLACE: $1$2
```

8 Isolating the major part-of-speech categories

Once an infrastructure of words with secure part-of-speech is in place, attention turns to attempts to broadly distinguish word classes.

8.1 Distinguishing verbs from nouns

The rules in this section aim to distinguish verbs from nouns.

(24). Isolating nouns that look like verbs by locating the heads of noun phrases

BACKGROUND: Some nouns happen to look like verbal forms. For example *bzaḥ* might be the future of *za* 'eat' or it might be a noun 'food'. The nominal reading is clear when the word heads a noun phrase, i.e. occurs before determiners and adjectives (e.g. *bzaḥ źim-po* 'tasty food').

RULE: If a word that has both [n.xxx] and [v.xxx] tags is followed by [d.xxx] or [adj] tags delete all of the [v.xxx] tags.

```
PATTERN:(\S+\|\S*\[n\.[^\]]*\]\S*?)(?:\[v\.[^\]]*\])+(\S*\s+\S+\|(?:\[(?:adj|d\.[^\]]*
)\])+\s+)
```

```
REPLACE: $1$2
```

(25). Isolating nouns that look like verbs by locating a preceding genitive

BACKGROUND: The preceding rule (24) made use of noun phrase structure to isolate nouns that head noun phrases from the verbs which they happen to resemble. Because it is only rule 40 that attempts to isolate the indefinite determiner *cig*, *źig*, *śig* from the imperative converb, which has homophonous forms, rule 24 is unable to use the indefinite determiner in its search for noun phrases, i.e. *gnas śig* is still ambiguous between 'a place' or 'reside!'. However, if a genitive precedes the word in question (e.g. *dben-paḥi gnas śig* a place which is isolated) then it is unambiguously a noun.

RULE: If a word has at least one hypothesized [v.xxx] tag and also has some other hypothesized tag, and this word comes after a word with a hypothesized [case.gen] tag, and comes before *źig*, *cig*, *śig*, then delete any [v.xxx] tags.

```
PATTERN:
(\S+\|\S*\[case\.gen\]\S*\s+\S+\|)(?:((?:\[(?!v\.)[^\]]*\])+)(?:\[v\.[^\]]*\])+|(?:\[v
\.[^\]]*\])+((?:\[(?!v\.)[^\]]*\])+))(\S*\s+(?:ཞིག|ཅིག|ཤིག)་?\|\S+)
```

```
REPLACE: $1$2$3$4
```

(26). Isolating relator nouns that look like verbs

BACKGROUND: Some forms, such as *skad*, can receive both relator noun [n.rel] (e.g. *ḥdi skad ces*) and verbal tags [v.invar] (e.g. *skad do*). Because the structure [case.gen] [n.rel] [case.xxx] is used to define relator nouns, the occurrence of a genitive to the left can be used to isolate secure relator nouns and deprecate verbal analyses.

RULE: If a word has [n.rel] and [v.xxx] as possible tags, and is preceded by something with the hypothesized tag [case.gen] then remove [v.xxx]

```
PATTERN: (\S+\|\S*\[case\.gen\]\S*\s+\S+\|\S*\[n\.rel\]\S*?)(?:\[v\.[^\]]*\])+(\S*)
REPLACE: $1$2
```

(27). Isolating nouns that happen to resemble imperative verbs
BACKGROUND: Some nouns, particularly *chos* 'dharma' happen to resemble imperative verbs. In this case *chos* 'prepare!' (pres. *ḥchos*). After the genitive case the nominal reading is likely and the imperative reading probably impossible.

RULE: If a word that follows [case.gen] has both the tags [n.count] and [v.imp] then the tag [v.imp] can be deleted.

```
PATTERN: (\S+\|\[case\.gen\]\s+\S+\|\S*\[n\.count\]\S*)\[v\.imp\](\S*)
REPLACE: $1$2
```

(28). Isolating numerals that happen to look like verbs
BACKGROUND: The syllable *bcu* can be both the future verb stem of the verb *ḥchu* 'draw water' and the cardinal number 'ten'. If this syllable occurs before a cardinal number it is very likely to also be a cardinal number.

RULE: If a word has both the tags [num.card] and [v.fut] and is followed by an unambigous cardinal number then delete from it the tag [v.fut].

```
PATTERN: (\S+\|\S*\[num\.card\]\S*)\[v\.fut\](\S*\s+\S+\|\[num\.card\]\s+)
REPLACE: $1$2
```

8.2 Disambiguating [neg] and [n.count]

Attention can now turn to tasks that rely on a distinction having been made, in so far as possible, between nouns and verbs. The interpretation of the words *mi* and *ma* as negation is only possible before verbs and verbal nouns. Consequently, it is only sensible to disambiguate the possible interpretations of *mi* and *ma* after a general attempt has been made to distinguish verbs and nouns.[8]

(29). Finding the nouns *mi and ma* within noun phrases
BACKGROUND: When the syllables *mi* or *ma* occur without a verb or verbal noun to their right, they cannot be negation. Conversely, if *mi* or *ma* occur followed by the end of a noun phrase, then they must be nouns. In many cases the presence of *mi* or *ma* within a noun phrase is signaled by the part-of-speech category of the following word.

8 It is not necessary to disambiguate *źig* [cv.imp] from *źig* [d.indef] (cf. rule 40) before disambiguating *mi* and *ma*, because the combination *mi źig* and *ma źig* are not ambiguous. Because [cv.sem] never comes after negation, there is no danger in tagging all *mi* before *źig* as [n.count]. In contrast, when we turn to disambiguate *źig* it will be helpful to already know that *mi* is a [n.count] because this will allow the disambiguation of *źig* in the context *mi źig* to [d.indef], without having to write any special rules.

At this point in the tagging the syllables *źig* and *ḥi* are not unambiguous (*źig* has the tags [d.indef] and [cv.imp]. *ḥi* has the tags [case.gen] and [cv.gen].), consequently it is not possible to specify them using their POS tags. Nonetheless, after either *ma* or *mi* these two syllables are unambiguously the end of a noun phrase. Concomitantly, the *ma* and *mi* must be within a noun phrase and can be tagged as nouns.

RULE:: If *mi* / *ma* is followed by an unambiguous [adj], [d.xxx], [n.count], [n.mass], [num.xxx], or [p.xxx], or by ambiguous *źig*, or *ḥi* then remove the [neg] tag.[9]

```
PATTERN: (((?:མི་|མ་)\|\S*\[n\.count\]\S*)\[neg\](\S*\s+)(\S+\|(?:\[(?:adj|d\.[^\]]*|n\.c
ount|n\.mass|num\.[^\]]*|p\.[^\]]*)\])+\s+|(?:ཞིག|འི)་?\|\S+)
```

```
REPLACE: $1$2$3
```

(30). Isolating *mi* [n.count] and *ma* [n.count] after the genitive

BACKGROUND: A genitive connects two nouns. Consequently, *mi* preceded by the genitive must either be a noun, or the first word of a noun phrase. In the former case *mi* can be tagged as a noun even if it precedes a present or future verb stem (e.g. *rmoṅ-pa ḥi mi ḥgroḥo* 'an ignorant person goes'). In the latter case, *mi* might still be negation (e.g. *bskal-pa graṅs med-pa ḥi mi dge-ba ḥi las* 'non virtuous deeds of countless eons'). It is important to isolate examples of the first type, because they would be otherwise be misanalysed as negation because of the following verb. In order to preclude the second type it suffices to specify that the word following *mi* is not a verbal noun.

No rule yet attempts to distinguish the genitive case from the genitive converb. Thus, in order to preclude the the morpheme preceeding *mi* is the genitive converb, it is necessary to add the stipulation that the word two before *mi* is not a verb stem.

The generalization that the genitive connects two nouns has one exception; the verb *rigs* 'to be proper' governs the genitive case. The syllable *mi* between a genitive and *rigs* is likely to be a negation marker (e.g. *rab tu ḥbyuṅ-ba ḥi mi rigs* 'it is not proper to take ordination'). Thus, the rule that uses a preceding genitive to locate instances of *mi* as a noun, must preclude that the following word is *rigs*.

A parallel argument applies to *ma*.

RULE: If *mi / ma* could be [n.count], follows a probable genitive, does not precede *rigs*, and does not precede a [n.v.xxx], and the word before the probable genitive is not an unambiguous [v.xxx] tag, then mark *mi / ma* as a [n.count].

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

```
REPLACE: $1$2
```

9 The caveat 'unambiguous' automatically excludes *dag* which can be both a verb and a plural suffix. The rule is written to specify [n.count] and [n.mass] only, because negation is perfectly permissible before [n.v.xxx].

(31). Isolating *mi* and *ma* as nouns before *med-pa*

BACKGROUND: In general if *mi* or *ma* precedes a verb, they are likely to be interpreted as negation. However, the negative verbal noun *med-pa* is already inherently negated, so if *mi* or *ma* precedes *med-pa* then they must be nouns rather than markers of negation.

RULE: If *mi / ma* precedes *med-pa* then assign the tag [n.count] to it.

```
PATTERN: ((?:^|\s)(?:མི་|མ་))\|\S+\s+(མེད་པ་?\|)
```

```
REPLACE: $1|[n.count] $2
```

(32). Identifying *mi/ma* [n.count] when conjoined by the associative case with an unambiguous noun

BACKGROUND: The associative case connects two nouns. Consequently, mi preceded by the associative must either be a noun, or the first word of a noun phrase. In the former case mi can be tagged as a noun even if it precedes a present or future verb stem (e.g. *lha daṅ mi ḥgroḥo* 'gods and people go'). In the latter case, mi might still be negation (e.g. *dge ba daṅ mi dge baḥi las* 'virtuous and non-virtuous deeds'). It is important to isolate examples of the first type, because they would be otherwise be misanalysed as negation because of the following verb. In order to preclude the second type it suffices to specify that the word following *mi* is not a verbal noun. An exception to the rule is made for the small number of verbs, such as *ldan* or *mthun*, which select for a noun phrase marked with associative case.

RULE: Tag *mi* or *ma* as a noun if it is conjoined by the associative case marker *daṅ* with an unambiguous noun (but not a verbal noun), unless *mi* or *ma* is followed by a verb (such as *ldan*,*mthun*, *bstun*, *phrad*, *mjal*, *ḥdra*, or *bcas*) which selects for a noun phrase marked with associative case.

```
PATTERN: (\S+\|(?:\[(?:d\.det|n\.(?!v)[^\]]*|num\.card|p\.pers)\])+\s+དང་\|\S*\s+(?:མི་|མ་)
\|)\S*(\[n\.count\])\S*(\s+)(?!(?:ལྡན་\||མཐུན་\||བསྟུན་\||ཕྲད་\||མཇལ་\||འདྲ་\||བཅས་\||\S+\|\S*\[n\.v\.[
^\]]*\]\S*))|((?:མི་|མ་)\|)\S*(\[n\.count\])\S*(\s+དང་\|\S*\s+\S+\|(?:\[(?:d\.det|n\.(?!v)[
^\]]*|num\.card|p\.pers)\])+\s+)
```

```
REPLACE: $1$2$3$4$5$6
```

(33). Identifying *mi* [neg] before present and future tense verb forms

BACKGROUND: If *mi* is followed by either present or future verbs, then it is probably [neg]. We must ignore the verb *gaṅ* since it can also be an interrogative pronoun (i.e. *mi gaṅ* can mean both 'not full' and 'which person'). The verb *sogs* 'etc.' and the verbal noun *bgres-pa* 'old', are also excluded; their semantics dictates that they are more likely to occur after the noun *mi* 'person' than they are to be negated. The inclusion of further text in the corpus will assuredly give rise to further occasions for ad hoc exceptions to this rule.

RULE: If *mi* which is ambiguous between [neg] and [n.count] is followed by a word (other than *gaṅ*, *sogs*, or *bgres-pa*) with the hypothesized tags [v.pres], [v.fut], [n.v.pres] or [n.v.fut] and could be [neg], then assign tag [neg] to *mi*.

```
PATTERN: ((?:^|\s)མི་)\|\S*\[neg\]\S*\s+(?!(?:གང་|སོགས་|བགྲེས་པ་)\|)(\S+\[n?\.?v\.(?:fut|pres)\]\S*)
```

```
REPLACE: $1|[neg] $2
```

(34). Identifying *ma* [neg] in the prohibitive
BACKGROUND: Although *ma* most characteristically negates the past, in the prohibitive construction it negates the present. This fact allows certain examples of *ma* to be securely analyzed as the negation prefix rather than the noun 'mother'.

RULE: If *ma* is followed by an unambiguous present verb stem, which in turn is followed by a possible imperative converb (i.e. *cig*, *źig*, *śig*), then assign [neg] to *ma*, and remove [d.indef] from *cig*, *źig*, *śig*.

```
PATTERN:
((?:^|\s)མ་)\|\S*\[neg\]\S*\s+(\S+\|\[v\.pres\]\s+\S*\|\S*\[cv\.imp\])(?:\[d\.indef\])?(\S*)
```

```
REPLACE: $1|[neg] $2$3
```

(35). Isolating *ma* [neg] before the past tense and *yin*
BACKGROUND: If *ma* is followed by past tense verbs or *yin*, then it is probably [neg]. The word 'mother' can occur in these positions, but its occurrence without any explicit nominal marking is likely to be exceedingly rare. It must be kept in mind nonetheless that this rule will yield some fals positives.

RULE: If *ma* is followed by an unambiguous present verb stem, which in turn is followed by a possible imperative converb (i.e. *cig*, *źig*, *śig*), then assign [neg] to *ma*, and remove [d.indef] from *cig*, *źig*, *śig*.
background: If *ma* is followed by past tense verbs or *yin*, then it is probably [neg]. The word 'mother' can occur in these positions, but its occurrence without any explicit nominal marking is likely to be exceedingly rare. It must be kept in mind nonetheless that this rule will yield some false positives.
RULE: If *ma* which is ambiguous between [neg] and [n.count] is followed by a word with the hypothesized tags [v.pres], [v.past], [n.v.pres], [n.v.past], or [v.cop] then assign tag [neg] to the word *ma*.

```
PATTERN: ((?:^|\s)མ་)\|\S*\[neg\]\S*(?=\s+(?:ཡིན་?\||\S+\[n?\.?v\.(?:past|pres)\]))
```

```
REPLACE: $1|[neg]
```

8.3 Isolating case markers and converbs

There is extensive overlap between the set of morphemes that serve as case marker and the set of morphemes that serve as converbs. In general, these morphemes are analyzed as case markers when they occur after noun phrases but are analyzed as converbs when they occur after verb stems. Because the distinction between case markers and converbs relies on the distinction between verbs and nouns, it is only possible to implement the rules in this section after the rules in section 8.1.

8.3.1 Disambiguating cases and converbs from other things

Before attempting to distinguish case markers and converbs from each other, we first distinguish case markers and converbs respectively from other things that they may happen to look like. In less abstract terms, section 8.3.2 provides rules that specify either [case.xxx] or [cv.xxx], but case and converbial markers suffer other types of ambiguity as well (e.g. ར་ <r(a)> can be [case.term], [cv.term] or [n.count], cf. rule 37). Such ambiguities should be resolved before the general question of case versus converb is addressed.

(36). Distinguishing nouns from cases and converbs at the left edge of noun phrases

BACKGROUND: There are syllables that are interpretable both as normal nouns and as morphological affixes (e.g. *nas* 'barely', *źiṅ* 'field', *las* 'deed', *śig* 'louse' versus *nas* elative case marker and elative converb, *źiṅ* imperfective converb, *las* ablative case marker and ablative converb, *śig* imperative converb and indefinite determiner). Because case markers and converbs must follow nouns and verbs respectively, at the left edge of noun phrases (i.e. after a *śad*, the genitive case or the associative case) only the noun interpretation is possible (e.g. ། *nas dkar mo* .. 'white barley', ། *źiṅ gi* 'of the field', *a-ma gaṅ ḥdod-pa ḥi las* 'whatever deed mother wishes').

RULE: If any word has at least two possible part-of-speech tags, one of them [n.count] and one more more that are either [cv.xxx] or [case.xxx], and this words appears directly after ། (or a *-g* without a *tsheg*), [case.gen] or [case.ass], then remove any tags [cv.xxx] and [case.xxx] tags from this word.

```
PATTERN: ((?:[།ག]\|\S+|\S+\|\[case\.(?:gen|ass)\])\s+\S+\|\S*?)(?:\[(?:case|cv)\.[^\]]*
\])+(\S*\[n\.count\]\S*)

REPLACE: $1$2
```

(37). Disambiguating <r(a)> ར་ as [n.count] and [case.term]/[cv.term]

BACKGROUND: The Tibetan syllable <r(a)> ར་ can be three things: the terminative case marker [case.term] after a noun phrase that ends in an open syllable (e.g. *rgyal-po r* 'to the king'), the terminative converb [cv.term] after a verb stem that ends in an open syllable (e.g. *za r ḥjug* 'make someone eat'), or the noun *ra* [n.count] 'goat'. However, the word *ra* 'goat' will have a *tsheg* that precedes it, but a *tsheg* will not precede the terminative case marker or terminative converb. At the very beginning of a sentence the noun ra 'goat' will not have a tsheg preceding it, but instead will have a *śad* or a *tsheg*-less final *ga* preceding it. An additional stipulation must be included in this rule because in the combinations *ga r* 'to where' and *dga r* 'to be happy' the letter 'ra' occurs with a preceding *tsheg*-less *ga*, but is nonetheless not the noun 'goat'.

37a. Identifying when <r(a)> ར་ is [n.count] rather than [case.term] or [cv.term]

RULE: If <r(a)> ར་ is preceded by a word that ends in ་, or by a sentence boundary (། or *tsheg*-less ག), then delete [case.term] and [cv.term] as analyses. An exception is made for preceding words ག or དག, which need not be sentence final.

```
PATTERN: ((?:་|།|(?<![\s་]ད?)ག)\|\S+\s+ར་\|\S*)\[case\.term\](\S*)\[cv\.term\](\S*)

REPLACE: $1$2$3
```

37b. Identifying when <r(a)> ར་ is [case.term] or [cv.term] rather than [n.count]
RULE: If <r(a)> ར་ which can still be [case.term] or [cv.term] comes after a word that does not end with ་, then delete [n.xxx] analyses from ར་, unless the preceding word is ག or ངག.

```
PATTERN: ((?:^|\s)(?!ང?ག\|)\S*[^་]\|\S+\s+ར་\|\S*\[(?:case|cv)\.term\]\S*?)(?:\[n\.[^\]]*
\])+(\S*)
```

```
REPLACE: $1$2
```

(38). Disambiguating -*s* ས་ the case suffix [case.agn] and *sa* ས་ 'earth' [n.count]
BACKGROUND: The letter <s(a)> ས་ can be the noun *sa* 'earth', the relator noun 'place' (*a-ma ḥi sa r* 'at mother's'), or the agentive case suffix *-s* after nouns that end with open syllables. In Tibetan orthography the case suffix ས་ is written together with the preceding syllable (e.g. *rgyal-po s* རྒྱལ་པོ ས་ 'king [case.agn] '). Consequently, *sa* ས་ 'earth' [n.noun] and 'place' [n.rel] can be differentiated from -*s* ས་ [case.agn] because *sa* ས་ 'earth' and 'place' [n.rel] are preceded by a word that ends in *tsheg*. At the very beginning of a sentence the noun *sa* will not have a *tsheg* preceding it, but instead will have a *śad* or a *tsheg*-less final *ga* preceding it.

38a. Identifying when -s ས་ is sa ས་ 'earth' [n.count] and not the case suffix [case.agn]
RULE: If <s(a)> ས་ is preceded by a word that ends in ་, or by a sentence boundary (། or tsheg-less ག), then delete [case.agn] as a possible analysis.

```
PATTERN: ((?:་|།|ག)\|\S+\s+ས་\|\S*)\[case\.agn\](\S*)
```

```
REPLACE: $1$2
```

38b. Identifying when -s ས་ is the case suffix [case.agn] and not sa ས་ 'earth' [n.count]
RULE: If <s(a)> ས་ which can still be [case.agn] comes after a word that does not end with ་, then delete [n.xxx] analyses from ས་.

```
PATTERN: (?<!་)(\|\S+\s+ས་?\|\S*\[case\.agn\]\S*?)(?:\[n\.[^\]]*\])+(\S*)
```

```
REPLACE: $1$2
```

(39). Distinguishing *de* [d.dem] from *de* [cv.sem]
It would be tempting to suggest that *de* [cv.sem] only comes after verbs, but this is incorrect. Although it is most frequent after verbs, the semifinal converb can follow almost any constituent. What can be said is that most *de* [d.dem] cannot occur after a verb stem, that most instances of the syllable *de* at the end of a noun phrase will be [d.dem], and that, because the semi-final converb ends a clause, there is a tendency for it to appear before a *śad*. These tendencies can be combined to isolate very likely instances of *de* [d.dem], namely those cases of *de* that occur at the end of a noun phrase (and thus not after a verb stem) and which are not followed by *śad.*

39a. Distinguishing *de* [d.dem] from *de* [cv.sem] in noun phrases
RULE: If *de* [d.dem] /[cv.sem] is preceded by a word with an unambiguous tag [adj], [d.xxx], [n.xxx], or [p.xxx], and is not followed by a *śad* then delete [cv.sem].

```
PATTERN:
(\S+\|(?:\[(?:adj|(?:d|n|num|p)\.[^\.\]]*)\])+\s+དེ་?\|\S*)\[cv\.sem\](\S*\[d\.dem\]\S*)
(?!\s+།)

REPLACE: $1$2
```

39b. Distinguishing *de* [d.dem] from *de* [cv.sem] at the end of a noun phrase

RULE:: If དེ is followed by ས or ར which can be case, then make ས or ར a case and remove [cv.sem] from དེ.

```
PATTERN: ((?:^|\s)དེ\|\S*)\[cv\.sem\](\S*\s+[སར]་?\|)\S*?((?:\[case\.[^\]]*\])+)\S*

REPLACE: $1$2$3
```

39c. Distinguishing *de* [d.dem] from *de* [cv.sem] after verb stems

RULE: If *de* [d.dem] /[cv.sem] is preceded by a word with an unambiguous tag [v.xxx] then delete [d.dem].

```
PATTERN: (\S+\|(?:\[v\.[^\]]*\])+\s+དེ་?\|\S*\[cv\.sem\]\S*)\[d\.dem\](\S*)

REPLACE: $1$2
```

(40). Distinguishing *cig*, *źig*, *śig* [cv.imp] from *cig*, *źig*, *śig* [d.indef] after the imperative and the prohibitive.

BACKGROUND: The syllable *cig* and its sandhi alternates *źig* and *śig* can either be an indefinite determiner (e.g. *lam cig* 'a path') or it can be a converb that marks the imperative (e.g. *khyed gñis kyis kho-bo sod cig* 'you two kill me!'). The imperative converb can only come after an imperative verb stem or a negated present verb stem in its prohibitive use (e.g. *grogs-po bdag ma gsod cig* 'O friends, do not kill me!), so in these context the interpretation as an indefinite determiner can be excluded. Conversely, outside of these two contexts the interpretation as an imperative converb can be excluded.

40a. Distinguishing *cig*, *źig*, *śig* [cv.imp] from *cig*, *źig*, *śig* [d.indef] after the imperative and the prohibitive

RULE: If any word has the two possible part-of-speech tags [cv.imp] and [d.indef], then delete the tag [d.indef] if the preceding word only has the tag [v.imp], or the preceding two words are *ma* and an unambiguous [v.pres].

```
PATTERN: ((?:^|\s)མ་\|\S+\s+\S+\|\[v\.pres\]|\S+\|\[v\.imp\])(\s+\S+\|\S*\[cv\.imp\]\S*
)\[d\.indef\](\S*)

REPLACE: $1$2$3
```

40b. Distinguishing *cig, źig, śig* [cv.imp] from *cig, źig, śig* [d.indef] elsewhere

RULE: If any word has the two possible part-of-speech tags [cv.imp] and [d.indef], and the preceding word has neither the tags [v.imp] or [v.pres], then delete the tag [cv.imp] from the word in question.

```
PATTERN: (\S+\|(?:\[(?!v\.(?:imp|pres))[^\]]*\])+\s+\S+\|\S*)\[cv\.imp\](\S*\[d\.indef
\]\S*)

REPLACE: $1$2
```

(41). Precluding *la* as a noun before the verb *thug*

BACKGROUND: The verb *thug* 'be at the point of' typically requires *la* as part of its rection. This *la* will be interpreted as a case marker after nouns and a converb after verbs, but it will never be interpretable as the noun 'mountain pass'.

RULE: If the syllable *la* precedes *thug* [v.xxx] or *thug-pa* [n.v.xxx], then remove from *la* the interpretation [n.count].

```
PATTERN: ((?:^|\s)ལ་\|\S*)\[n\.count\](\S*\s+ཐུག(?:་པ)?་?\|\S+)

REPLACE: $1$2
```

(42). Precluding *la* as a noun in clause final position

BACKGROUND: The syllable *la* has many interpretations: the allative case, the allative converb, the stem of the proclausal adverb *lar* 'moreover', and the noun mountain pass. At the end of a clause (i.e. after a verb or verbal noun but before a *śad*) the noun 'mountain pass' can be precluded.

RULE: If a word *la* appears after [v.xxx] or [n.v.xxx] and before །, then delete [n.count] from this *la*.

```
PATTERN: (\S+\|(?:\[(?:n\.)?v\.[^\]]*\])+\s+ལ་?\|\S*)\[n\.count\](\S*\s+།\|\S+)

REPLACE: $1$2
```

(43). Precluding *nas* as a noun in clause final position

BACKGROUND: The syllable *nas* has many interpretations: the ellative case, the ellative converb, and the noun 'barley'. At the end of a clause (i.e. after a verb but before a *śad*) the noun 'barley' can be precluded.

RULE: If a word *nas* appears after [v.xxx] and before །, then delete [n.count] from this *nas*.

```
PATTERN: (\S+\|(?:\[v\.[^\]]*\])+\s+ནས་\|\S*)\[n\.count\](\S*\s+།\|\S+)

REPLACE: $1$2
```

8.3.2 Distinguishing cases and converbs from each other

A case marker is affixed to a noun phrase and a converb is affixed directly to a verb stem. A nominalized verb counts (for most purposes) as a noun. There are two exceptions to the overall pattern of cases after noun phrases and converbs after verb stems. We allow the locative converb after a verbal noun when there is a clear converbial meaning 'when/if' rather than 'in', the typical case meaning (e.g. *mi ḥi naṅ du skyes-pa na / … gcig la gcig ḥtshe źiṅ gnod-pa r gyur to* । 'When born among men … they hurt and harm one another'.

We rarely analyse the genitive case marker as appearing directly appended to a verb stem. For example, *ḥbaṅs ḥdi dag thams-cad la mgon-skyabs daṅ | gnas med-par ḥgyur gyi mi ḥgaḥ tsam gyi phyi r |* ‘all these subjects will be some mere men without a protector or place’ and *soṅ gi phyir* in the sentence *mdaḥ gźu blaṅs nas rgyal-po ñid lag dar te khyeḥu la ḥphaṅs nas mdaḥ ḥphangs pa khyeḥu lam soṅ gi phyir yaṅ rgyal-poḥi druṅ du lhuṅ ṅo* ‘Taking up a bow and arrow, the king himself drew back his hand and shot at the person. The arrow that he shot after the path the person had taken landed in front of the king’ (cf. Garrett et al. forthcoming).

(44). Isolating case markers after nominals

BACKGROUND: When the element to the left of a syllable that can be either a case marker or converb is unambiguously part of a noun phrase, interpretation of the syllable as a converb can be excluded. This rule must be implemented in three stages. In the first stage, converbial interpretations are excluded after elements of noun phrases in general.

However, because *de* [d.dem] and *cig, źig, śig* [d.indef] are not yet distinguished from the homophonous *de* [cv.sem] and *cig, źig, śig* [cv.imp], it is not possible to locate case markers after them using a search for the tags [d.dem] and [cv.sem]. Instead, a second stage of the rule takes aim at the phonological material of these morphemes, paying no attention to their interpretation. This strategy is safe, because combinations such as *de la* or *cig gi* are securely interpretable respectively as the demonstrative in the allative case and an indefinite marker in the genitive case.

Because we permit [cv.loc] after verbal nouns the most general form of this rule must allow converbs after verbal nouns. Consequently, a second rule narrows in specifically on verbal nouns followed by converbs other than [cv.loc].

44a. Isolating case markers after nominals other than verbal nouns, [d.dem] and [d.indef]

RULE: If homophonous [case.xxx]/[cv.xxx] is preceded by a word with an ambiguous tag *de*, *cig*, *źig*, or *śig*, or an unambiguous tag [adj], [d.xxx], [n.count], [n.mass], [n.rel], [num.xxx], or [p.xxx], then delete the [cv] tag.

```
PATTERN: ((?:^|\s)(?:དེ་|ཅིག་|ཞིག་|ཤིག་)\|\S+|\S+\|(?:\[(?:adj|(?:d|n|num|p)\.[^\.\]]*)\])+)(\s
+\S+\|\S*?(?:\[case\.[^\]]*\])+\S*?)(?:\[cv\.[^\]]*\])+(\S*)

REPLACE: $1$2$3
```

44b. Isolating case markers after verbal nouns

RULE: If homophonous [case.xxx] /[cv.xxx], which is not *na* [case.loc]/[cv.loc] is preceded by a word with an unambiguous tag [n.v.xxx] then delete the [cv] tag.

```
PATTERN:
(\S+\|(?:\[n\.v\.[^\]]+\])+\s+\S+\|\S*\[case\.(?!loc)[^\]]*\]\S*)\[cv\.[^\]]*\](\S*)

REPLACE: $1$2
```

(45). Isolating converbs after verbs

BACKGROUND: Now we turn from isolating secure instances of [case.xxx], to isolating secure instances of [cv.xxx]. After unambiguous verb stems, morphemes that are ambiguously case markers or converbs (other than the genitive) can be specified as converbs.

RULE: If a word with the hypothesized tags [case.xxx] ~ [cv.xxx] directly follows a word that is only tagged with [v.xxx] then the tag [case.xxx] can be removed, n.b. except that we do not automatically remove [case.gen], because it is permitted after verb stems.

```
PATTERN:
(\S+\|(?:\[v\.[^\]]*\])+\s+\S+\|\S*)\[case\.(?!gen)[^\]]*\](\S*\[cv\.[^\]]*\]\S*)

REPLACE: $1$2
```

(46). Specifying *tu* and *du* as converbs after *śin*

BACKGROUND: A specific rule is necessary to treat *śin-tu*. We treat *śin-tu* as an infinitive construction, although *śin* is not otherwise attested as a verb, which is why it is not tagged like one. In our system *tu* and *du* are to be tagged as converbs after *śin*.

RULE: If *du* or *tu* follows *śin* [adv.intense] then [case.term] can be deleted as an option.

```
PATTERN: ((?:^|\s)ཤིན་\|\S*\[adv\.intense\]\S*\s+(?:ཏུ|དུ)་?\|\S*)\[case\.term\](\S*)

REPLACE: $1$2
```

(47). Specifying *la* as a case marker in the phrase *la sogs-pa*

BACKGROUND: In the preceding rules the unambiguous right edges of noun phrases and unambiguous verb stems to the left of the [case.xxx]/[cv.xxx] permitted disambiguation. An alternative approach is to look to the right of the [case.xxx]/[cv.xxx]. If to the right of an ambiguous [case.xxx]/[cv.xxx] is a verb which requires that particular case in its rection, then the sequence can be assigned the tag [case.xxx]. So far we have only one rule of this type. Etymologically the phrase *la sogs-pa* 'etc.' is a case marker followed by a verbal noun 'gathered at'. This analysis is clear in the Old Tibetan spelling *la stsogs pa*. In general our tact is to err in favor of etymologically faithful analyses, in the absence of compelling evidence to the contrary. Consequently, the *la* in the phrase *la sogs-pa* can be specified as a case marker.

RULE: If *la* is followed by *sogs-pa* then assign [case.all] to *la* (i.e. remove other possible tags, [cv.all] and [n.count]).

```
PATTERN: ((?:^|\s)ལ་)\|\S+\s+(སོགས་པ་?\|)

REPLACE: $1|[case.all] $2
```

One could also introduce a rule to assign *daṅ* the tag [case.ass] before the verb *mjal*. But the data in our training corpus has not yet prompted such a rule. Further research into Tibetan case rection would doubtless give rise to additional such rules.

9 Distinguishing types of nominals

The part-of-speech tag set does not distinguish very many types of nouns. The rules in this section seek to find syntactic patters that permit the isolation of one type of noun from another.

9.1 Distinguishing nouns from relator nouns

It frequently happens that a relator noun coincides with a lexical noun; this reflects the origin of most relator nouns as grammaticalized nouns. For example, *nań* can mean both ‘the inside’ (*'bras-bu phyi-rol smin la* । *nań ma smin-par* ‘the outside of the fruit was ripe, but it’s inside was not ripe…), but also mean ‘inside (of)’ (*me-loń gi nań du* ‘inside the mirror’ …).

(48). Isolating relator nouns after a genitive and before a spatial case

BACKGROUND: Garrett et al. (forthcoming) define a relator noun as having “a genitive before it and a spatial case (allative, locative, terminative) after it”. The tagger may consequently use the same syntactic frame to confidently isolate relator nouns.

RULE: If word has two possible tags [n.count] and [n.rel] and it occurs after a possible [case.gen] and before a [case.term], [case.loc], or [case.all], then delete the tag [n.count].

```
PATTERN: (\S+\|\S*\[case\.gen\]\S*\s+\S+\|\S*)\[n\.count\](\[n\.rel\]\S*\s+\S+\|\[case
\.(?:all|loc|term)\])
```

```
REPLACE: $1$2
```

(49). Isolating nouns in clause initial position

BACKGROUND: Relator nouns relate a constituent on the right to a constituent on the left. Consequently, if there is no constituent to the left of a word it is unlikely that this word is a relator noun.

RULE: If word has two possible tags [n.count] and [n.rel] and it occurs after a *śad* (or a *-g* not followed by a *tsheg*), then delete the tag [n.rel].

```
PATTERN: (\S*(?:ག|།)\|\S+\s+\S+\|\S*\[n\.count\])\[n\.rel\](\S*)
```

```
REPLACE: $1$2
```

9.2 Isolating reflexive pronouns

(50). Isolating *rań* as a reflexive pronoun

BACKGROUND: The syllable *rań* is analyzable both as a reflexive pronoun (*ńed rań* ‘we ourselves’, *khyed rań* ‘you yourselves’, *a-ma na-re rań gi nor la* ‘Mother said “for one’s own wealth…”’) and as a determiner (*że-sdań chen-po rań cig* ‘a very great antipathy’). After a personal pronoun the determiner use can be excluded.

RULE: If syllable *rań* occurs after a word with the tag [p.pers], then delete from *rań* the analysis [d.det].

```
PATTERN: (\S+\|\[p\.pers\]\s+རང་\|\S*)\[d\.det\](\S*)
```

```
REPLACE: $1$2
```

9.3 Isolating names

Named entity recognition is a challenging area of natural language processing, which largely falls outside of the scope of our project. Nonetheless, names are typically introduced in a text using fixed constructions. These fixed constructions permit the identification of previously unknown words as names.

(51). Identifying unknown words as names

BACKGROUND: It is very common in Tibetan texts that the first time a protagonist is introduced by name, his name will appear before *źes bya-ba;* this fact allows words of unknown meaning to be interpreted as names in this context.

RULE: If a word without any assigned analysis immediately precedes *źes* [cl.quot] *bya-ba* [n.v.fut] then assign this word the tag [n.prop].

```
PATTERN: ((?:^|\s)[^\[]+)\s+(ཞེས་\|\[cl\.quot\]\s+བྱ་བ་?\|\[n\.v\.fut\]\s+)
```

```
REPLACE: $1|[n.prop] $2
```

10 Distinguishing the four tenses and subsequent cleanup

The most tricky aspect of assigning part of speech tags yet confronted in our project is the disambiguation of verb tense. The Tibetan verbal system is not well understood. As a default hypothesis, we follow the dictionaries (cf. Hill 2010) in assuming that all Tibetan verbs in principle distinguish four tenses, the present [v.pres], the past [v.past], the future [v.fut], and the imperative [v.imp]. If certain verbs lack an imperative, as for example the *Tshig mdzod chen mo* (Zhang 1989) and the *Dag yig gsar bsgrig* (Bsam gtan 1979) believe, this fact will emerge from the corpus; it is not to be written into its architecture.

10.1 Disambiguating verb tenses

When all four tenses of a verb are morphologically distinct (e.g. pres. *gsod*, past *bsad*, fut. *gsad*, imp. *sod*) the lexicon alone succeeds at disambiguating one from the other. However, when the stems are partially or entirely ambiguous, one must seek other means to disambiguate one stem from another. Two such means are in general available.

First, certain syntactic contexts only permit certain tenses (cf. §10.1.3 and §10.1.4), viz. *ma* only negates the past and the present in its prohibitive use; *mi* never negates the past; the future never appears before *nas*. Second, certain sandhi contexts imply the presence (or absence) of the *da-drag* (§10.1.5), which typically is a marker of the past.

10.1.1 The correct ordering of disambiguation strategies

In a sense, if there is evidence for a *da-drag*, then the stem is not actually ambiguous (i.e. *gsol to* = *gsold to* versus *gsol lo*). Consequently, one might think that it is preferable to run the *da-drag* rules before the syntactic disambiguation rules, with the intuition that brute facts before the eyes should take epistemological preference over syntactic implications. In an earlier version of these rules, we followed precisely this course. However, in our experience syntactic contexts are more reliable than the presence of a *da-drag*.

Consider the example which shows the need to run the rule which forbids the part-of-speech sequence [v.imp] [cv.ques], before running the *da-drag* detection rules. *gaṅ źig śin tu dad-pa ḥi sems kyis chu sñam-pa gaṅ tsam saṅs-rgyas la mchod dam* | *dge ḥdun la phul* [v.past] ~ [v.imp] *lam* | *pha-ma la phul* [v.past] ~ [v.imp] *lam* | *dbul-phoṅs la byin nas* | *gcan-gzan la byin na* | *gsod-nams ḥdi ni bskal-ba du-ma r yaṅ mi zad de* | "If one with a mind of great faith offers handfuls of water to the buddha, or makes offerings to the saṃgha, or makes offerings to one's parents, or gives to the poor, or gives to wild animals, this merit shall not run out for many eons". Here the verb stem *phul* is morphologically either past or imperative (pres. *ḥbul*, past *phul*, fut. *dbul*, imp. *phul*). If the rule based tagger searches for the *da-drag* before taking into account syntactic disambiguation, the analysis [v.past] is removed because the form of the question converb is *lam* and not *tam*. After removing the [v.past] analysis the only the analysis [v.imp] remains. Consequently, the analysis [v.imp] is not removed by the rule that forbids a [v.imp] before a [cv.ques]. The analysis [v.past] is intelligible in this example, but the analysis [v.imp] makes no sense at all.[10]

Another example shows the need to use negation to disambiguate verb stems before running the *da-drag* rules. In some cases, because of what we might want to call 'errors' in the text, different cues point toward divergent analyses. For example, *ku-śu ḥdi ni ḥbras-bu las skyes-pa ma lags te* | *chab-mig cig gi naṅ nas rñed-pa s slan-cad ni bdag gis mi rñed de* | *mi ḥbyor* v.past v.pres *to* "This apple was not born from fruit, but I found it from inside a spring, so I cannot find it hereafter. It will not be encountered." In this example the negation with *mi* suggests a present or future verb stem, but the form of the final converb *to*, by implying the presence of a *da-drag* suggests that v.past is the correct analysis of the verb stem.

Here is a similar example, *gal te sñiṅ nas ma btsal* v.past v.fut *lam le-lo źig byas te* | … 'if one does not seek wholeheartedly, or is lazy… '. The negation with *ma* suggests that *btsal* should be analyzed [v.past], but the form of the question converb *lam* rather than *tam* suggests that [v.fut] is the correct analysis. The coordination of *btsal* with *byas* however confirms that negation should be trusted as the correct cue, and that the absence of the *da-drag* here points in the wrong direction. The dictionaries do suggest that *btsal* should have a *da-drag* in the past (cf. Hill 2010: 242).

These examples make clear that it is necessary to run the syntactic disambiguation rules before the *da-drag* rules.

In view of the fact that the *da-drag* was moribund by the time of Classical Tibetan, and indeed its use in Old Tibetan is not fully understood, we take the cues provided by negation as more persuasive than those provided by the evidence for the *da-drag*. It is clear however, that these cases of conflicting evidence for the interpretation of the verb stem deserve to be studied more systematically as they may reveal a great deal about the Tibetan verbal system, or at least, the development of editorial practices.

Whereas the rules that used sandhi contexts to disambiguate verb stems, were relying primarily on phonological evidence, so that in a sense the stems themselves are not ambiguous across the paradigm, the following rules use syntactic cues alone to disambiguate verb stems.

10 It is noteworthy that none of the dictionaries in fact gives *phuld*, but only *phul*, as the past of this verb (cf. Hill 2010: 204).

10.1.2 Isolating auxiliary verbs

(52). Isolating auxiliary verbs

BACKGROUND: In Canonical Tibetan a limited number of verbs occur as auxiliary verbs (*nus* 'be able', *dgos* 'need', *śes* 'know', *ran* 'be time for', *srid* 'be possible'). These auxiliaries come directly after the main verb of a clause, except for the possible interposition of a negation marker. This distribution allows these auxiliaries to be easily identified. It is important to isolate auxiliaries before running the tense disambiguation rules, because otherwise auxiliaries would have to be written in as exceptions to some of these rules.

RULE: If a word with the possible analysis [v.aux] either (1) follows a word that only has verb stem analyses, i.e. [v.xxx], or (2) follows a sequence of such a word and a negation prefix, i.e. [v.xxx] [neg], then retain [v.aux] as the only possible [v.xxx] analysis for the word.

```
PATTERN: (\S+\|(?:\[v\.[^\]]*\])+(?:\s+\S+\|\[neg\])?\s+\S+\|\S*\[v\.aux\])(?:\[v\.[^\
]]*\])+(\S*)

REPLACE: $1$2
```

10.1.3 Using co-occurrence with converbs to disambiguate verb tenses

In certain tenses verb stems are incompatible with certain converbs. The imperative is quite restricted in its distribution. The circumscribed syntactic occurrence of the imperative allows us to specify a number of situations in which it can be securely located, and other circumstances where the analysis of a word as an imperative is impossible. The imperative converb only follows imperative tense verb stems. The imperative does not occur in subordinate clauses, so converbs that imply subordination, such as the semi-final converb, preclude the interpretation of the preceding stem as an imperative. In finite contexts the final converb and question converb make clear that a sentence is not imperative. Rules 53-55 take advantage of restrictions on the imperative to disambiguate verb tenses. We know of fewer co-occurrence restrictions for the other tenses. The future does not occur before the elative converb, a fact that rule 56 takes advantage of.

(53). Finding the imperative before [cv.imp]

BACKGROUND: If an ambiguous imperative verb stem occurs before an ambiguous imperative converb (e.g. *gśegs śig* 'go!'), then the analysis as an imperative verb stem and an imperative converb is secure. Two possible exceptions occur. 1. The imperative converb follows a negated present in the prohibitive (e.g. *ma gśegs śig*). Consequently, the rule must stipulate that a *ma* does not precede the ambiguous verb stem. 2. If the imperative verb stem can also be a noun, since the imperative converb can also be an indefinite determiner the phrase is ambiguous (e.g. *gnas śig* 'stay!' or 'a place'). However, this exception need not be a cause for concern so long as the rule only removed hypothesized tenses other than the imperative, rather than stipulating interpretation as the imperative.

We turn now from the imperative to the future. The future verb stem does not occur before *nas* [cv.ela]; this gap in its distribution allows us to disambiguate the tense of many verb forms.

RULE: If a word with the hypothesized part-of-speech-tag [v.imp] is followed by *cig*, *źig*, or *śig* (and is not preceded by *ma*) then delete all other hypothesized [v.xxx] tags.

```
PATTERN:
((?:^|\s)(?!མ་\|)\S+\s+\S+\|\S*?)(?:\[v\.[^\]]*\])*(\[v\.imp\])(?:\[v\.[^\]]*\])*(\S*\s
+(?:ཅིག|ཞིག|ཤིག)་?\|\S+)

REPLACE: $1$2$3
```

(54). Finding the prohibitive (present negated with *ma*) before [cv.imp]

BACKGROUND: The imperative converb follows a negated present in the prohibitive (e.g. *ma gśegs śig*). Consequently, an ambiguous verb stem can be stipulated as present in this circumstance.

RULE: If a word tagged with a hypothesized part-of-speech-tag [v.pres] is followed by *cig*, *źig*, or *śig* and is preceded by *ma* then delete all other hypothesized part-of-speech-tags.[11]

```
PATTERN: ((?:^|\s)མ་\|\S+\s+\S+\|)\S*(\[v\.pres\])\S*(\s+(?:ཅིག|ཞིག|ཤིག)་?\|)

REPLACE: $1$2$3
```

(55). Prohibiting the imperative in non-finite and finite but explicitly non-imperative contexts

BACKGROUND: The imperative is generally not permitted before converbs, or other non-finite contexts (such as before *kyaṅ*). It is likely that further training data will prompt the inclusion of further contexts in which the imperative is impossible.

RULE: If a word has more than one [v.xxx] tag, including [v.imp], and the following word either has the form *na*, *kyaṅ*, *yaṅ*, *nas*, or has any of the tags [cv.ela], [cv.fin], [cv.impf], [cv.loc], [cv.ques], [cv.sem], or [cv.term], then remove the tag [v.imp] from the word in question.

```
PATTERN:
(\S+\|\S*)(?:(\[v\.[^\]]*\])\[v\.imp\]|\[v\.imp\](\[v\.[^\]]*\]))(\S*\s+(?:(?:ན|ཀྱང|ཡང|ནས
)་?\|\S+|\S+\|\S*\[cv\.(?:ela|fin|impf|loc|ques|sem|term)\]\S*))

REPLACE: $1$2$3$4
```

(56). The prohibition of the future before the elative converb *nas*

BACKGROUND: The future tense verb stem does not occur before the elative converb *-nas*. Consequently, if an ambiguous verb stem occurs before the elative converb, the interpretation of the verb in question as a future can be precluded.

RULE: If a word has more than one [v.xxx] tag, including [v.fut] and the following word is *nas*, remove the tag [v.fut] from the word in question.

```
PATTERN:
(\S+\|\S*)(?:(\[v\.[^\]]*\])\[v\.fut\]|\[v\.fut\](\[v\.[^\]]*\]))(\S*\s+ནས་?\|\S*\[cv\.
ela\]\S*)

REPLACE: $1$2$3$4
```

11 Rule 40b has already stipulated that *cig*, *źig*, and *śig* are tagged as [cv.imp] in this context.

10.1.4 Using negation to disambiguate verb stems

The restriction of negation to certain verb stems is useful for disambiguation. Negation with *ma* occurs only with the past (and with the present in its prohibitive function, dealt with above in rule 54). Negation with *mi* precludes the past, but is possible with both the present and future.

(57). Isolating verb stems and verbal nouns after negation with *ma*

BACKGROUND: Negation with ma occurs with the past, and with copulas and auxiliary verbs, which our system does not distinguish for tense. (It can also occur with the present in its prohibitive function, which was dealt with above in rule 54.) Therefore, if a reasonable effort has already been made to isolate those cases of *ma* that are the noun 'mother' (rules 29-35), where possible, it is safe to assume that only these stem forms, or their nominalized equivalents, can follow negation with *ma*.

RULE: If a word tagged with a hypothesized part-of-speech-tag [v.aux] ([n.v.aux]), [v.cop] ([n.v.cop]), or [v.past] ([n.v.past]) is preceded by ma [neg], then delete all other hypothesized tags.

```
PATTERN: (མ་\|\[neg\]\s+\S+\|)\S*?(?:(\[(?:n\.)?v\.(?:aux|cop)\])\S*(\[(?:n\.)?v\.past\
])|(\[(?:n\.)?v\.(?:aux|cop)\])\S*|(\[(?:n\.)?v\.past\]))\S*

REPLACE: $1$2$3$4$5
```

(58). Precluding the past after negation with *mi*

BACKGROUND: Negation with *mi* precludes the past, but it possible with both the present and future.

RULE: After *mi* [neg], keep only [v.aux], [v.fut], [v.pres], [n.v.aux], [n.v.fut], and [n.v.pres].

```
PATTERN: (མི་\|\[neg\]\s+\S+\|)(?:\[(?!(?:n\.)?v\.)[^\]]*\])*(\[(?:n\.)?v\.aux\])?(?:\[(
?:n\.)?v\.cop\])?(\[(?:n\.)?v\.fut\])?(?:\[(?:n\.)?v\.imp\])?(?:\[(?:n\.)?v\.neg\])?(?
:\[(?:n\.)?v\.past\])?(\[(?:n\.)?v\.pres\])?\S*

REPLACE: $1$2$3$4
```

10.1.5 Using the presence (or absence) of a *da-drag* to disambiguate verb stems

In principle the (orthographically) lost *da-drag* helps to distinguish the stems of many verbs whose stems end in *-r*, *-n*, and *-l*. Although not normally written in Classical Tibetan, the *da-drag* makes itself known through its sandhi effects. In particular, the allomorphs *to* [cv.fin], *kyaṅ* [cl.focus], *ciṅ* [cv.impf], and *tam* [cv.ques] when occurring verb stems ending in *-r*, *-n*, and *-l* make clear that the verb stem in question is has a *da-drag* (i.e. is [v.past] and not [v.fut]). Conversely, verb stems ending in *-r*, *-n*, and *-l*, which are followed by other allomorphs of these morphemes, do not have the *da-drag* and thus the [v.past] reading can be excluded.

(59). The *da-drag* before *kyaṅ*, *ciṅ*, *to*, *tu*, or *tam*

BACKGROUND: A final *da-drag* is typical of past verbs with roots that end in *-n*, *-r*, *-l* (e.g. pres. *sbyin*, past *byind* 'give'); the *da-drag* can however also occur as the final of presents (e.g. pres. *seld*, past *bsald* 'cleanse, remove'). The presence of a *da-drag* has ramifications on the sandhi determined allomorphs of the following word in a number of cases. Specifically, after a *da-drag* one sees *kyaṅ* [cl.focus], *ciṅ* [cv.impf], *to* [cv.fin], *tu* [cv.term], and *tam* [cv.ques] rather that other allomorphs of these morphemes, such as *yaṅ*, *źiṅ*, *no* etc., *du*, and *nam* etc.

Not all past form that could have a *da-drag* do have a *da-drag*. For example, the verb 'give' (*sbyin, byin, sbyin, byin*) appears with the final converb as *byin no* and not *byin to*. Consequently, when there is no morphological ambiguity among present, past, and future, it would be inappropriate to insist on, or indeed expect, a *da-drag*. The temptation looms to only make use of *da-drag* information when a stem is ambiguous between past and future, but such a specification of the rule also has disadvantages. The rule of sections 10.1.3 and 10.1.4 will have already deleted many analyses from verbs that are in principle ambiguous between past and future. This action would delete the trigger for a rule that requires ambiguity between the past and future. The solution we have achieved is to insist only that a verb stem is somehow still ambiguous. It would be senseless to delete [v.past] from *byin* if it is the only remaining verbal analysis.

RULE: If a word has more than one [v.xxx] tag, including [v.fut], and the word ends in -l, -n, or -r and is followed by the word *kyaṅ*, *ciṅ*, *to*, *tu*, or *tam* then delete [v.fut].

```
PATTERN:
(\S+[ནརལ]་\|\S*)(?:(\[v\.[^\]]+\])\[v\.fut\]|\[v\.fut\](\[v\.[^\]]+\]))(\S*\s+(?:ཀྱང|ཅིང|ཏོ|ཏུ
|ཏམ)་?\|)

REPLACE: $1$2$3$4
```

(60). The absence of the *da-drag* before the final converb

BACKGROUND: After a *da-drag* the final converb takes the form *to*. Consequently, the forms of the final converb *no, ro,* and *lo* can be taken as evidence for the absence of a *da-drag*, which in turn provides evidence against the interpretation of the verb in question as a past.

RULE: If a word has more than one [v.xxx] tag, including [v.past], and this word ends in -l, -n, or -r and is followed by the word *no, ro*, or *lo*, then delete [v.past].

```
PATTERN:
(\S+([ནརལ])་\|\S*)(?:(\[v\.[^\]]+\])\[v\.past\]|\[v\.past\](\[v\.[^\]]+\]))(\S*\s+\2\u0F
7C་?\|)

REPLACE: $1$3$4$5
```

(61). The absence of the *da-drag* before *źiṅ*

BACKGROUND: After a *da-drag* the imperfective converb takes the form *ciṅ*. Consequently, the form *źiṅ* of the imperfective converb can be taken as evidence for the absence of a *da-drag*, which in turn provides evidence against the interpretation of the verb in question as a past.

RULE: If a word has more than one [v.xxx] tag, including [v.past], and this word ends in -l, -n, or -r and is followed by the word *źiṅ* then delete [v.past].

```
PATTERN:
(\S+[ནརལ]་\|\S*)(?:(\[v\.[^\]]+\])\[v\.past\]|\[v\.past\](\[v\.[^\]]+\]))(\S*\s+ཞིང་\|)
```

```
REPLACE: $1$2$3$4
```

(62). The absence of the *da-drag* before [cv.ques]

BACKGROUND: After a *da-drag* the question converb takes the form *tam*. Consequently, the forms of the question converb *nam, ram,* and *lam* can be taken as evidence for the absence of a *da-drag*, which in turn provides evidence against the interpretation of the verb in question as a past.

RULE: If a word has more than one [v.xxx] tag, including [v.past], and this word ends in -l, -n, or -r and is followed by *lam* [cv.ques] *nam* [cv.ques] or *ram* [cv.ques], then remove the tag [v.past].

```
PATTERN:
(\S+([ལནར])་\|\S*)(?:(\[v\.[^\]]+\])\[v\.past\]|\[v\.past\](\[v\.[^\]]+\]))(\S*\s+\2མ་?\|
\[cv\.ques\])

REPLACE: $1$3$4$5
```

(63). The absence of the *da-drag* before [cv.term]

BACKGROUND: After a *da-drag* the terminative converb takes the form *tu*. Consequently, the form *du* of the terminative converb can be taken as evidence for the absence of a *da-drag*, which in turn provides evidence against the interpretation of the verb in question as a past.

RULE: If a word has more than one [v.xxx] tag, including [v.past], and this word ends in -l, -n, or -r and is followed by *du* then remove the tag [v.past].

```
PATTERN:
(\S+[ལནར]་\|\S*)(?:(\[v\.[^\]]+\])\[v\.past\]|\[v\.past\](\[v\.[^\]]+\]))(\S*\s+དུ་?\|)

REPLACE: $1$2$3$4
```

(64). Removing the future for verbal nouns ending in *-pa*

BACKGROUND: Inside of verbal nouns the *da-drag* can also be detected. The nominalization suffix takes the form *-ba* after *-r*, *-n*, or *-l*, but is *-pa* after the *da-drag*, i.e. implying that the verb stem is [n.v.past].[12]

RULE: If a word has more than one [n.v.xxx] tag, including [n.v.fut], and the verb stem ends in -l, -n, or -r and is nominalized with *-pa*, then remove the tag [n.v.fut] from this word.

```
PATTERN:
(\S+[ལནར]་པ་?\|\S*)(?:(\[n\.v\.[^\]]+\])\[n\.v\.fut\]|\[n\.v\.fut\](\[n\.v\.[^\]]+\]))(\S
*)

REPLACE: $1$2$3$4
```

---

12 Because *-pa* and *-ba* are similar looking and frequently confused, this rule may seem to risk introducing errors. However, we think it is best to disambiguate verb stems wherever it is possible to do so. Disambiguating these stems permits the behavior of *-pa* versus *-ba* to be more easily explored by future researchers; reason enough to add the rule.

(65). Removing the past for verbal nouns ending in *-ba*

RULE: If a word has more than one [n.v.xxx] tag, including [n.v.past], and the verb stem ends in -l, -n, or -r and is nominalized with *-ba*, then remove the tag [n.v.past] from this word.

```
PATTERN:
(\S+[ལནར]་བ་?\|\S*)(?:(\[n\.v\.[^\]]+\])\[n\.v\.past\]|\[n\.v\.past\](\[n\.v\.[^\]]+\]))(
\S*)

REPLACE: $1$2$3$4
```

10.2 Consolidating ambiguous verbs forms into ambiguous tags

In many cases it is not possible to decide which tense of a verb is being used in a given situation. One would like to leave these cases ambiguous, however, the proofreading capacity of the rule tagger is only triggered if it achieves an unambiguous analysis of a given word. Consequently, it is advantageous to replace ambivalent tagging with unequivocal tagging of a verb stem as ambiguous. The tags [v.fut.v.pres], [v.past.v.fut], [v.fut.v.past], and [v.invar] (which could equivalently have been [v.fut.v.past.v.pres]), allow unambiguously ambiguous verb stems. A parallel suite of tags is also added for the verbal nouns (cf. §4).

There is a substantial disadvantage to introducing these tags. Without them the pre-tagger would have led to suggestions such as *bskyed* [v.fut] ~ [v.past]. Now these will be changed to *bskyed* [v.fut.v.past], but once the lexicon is recompiled the *bskyed* [v.fut.v.past] will move from the training data into the lexicon. After this, the pre-tagger would suggest *bskyed* [v.fut] ~ [v.past] ~ [v.fut.v.past], because the system has no way to know that the suggestion [v.fut.v.past] adds no information. We avoid these complications by taking apart these ambiguous verb stems into their constituent parts, as the very first step of the rule based tagging (cf. rules 1-4).

(66). The creation of the tags [v.invar] and [n.v.invar]

BACKGROUND: In many syntactic contexts the rule based tagger will be unable to unambiguously specify the choice of verb stems. For example, *gśegs so* could be present, past, or future. Because such contexts are systematically ambiguous it would not be useful to present the human user with a choice between these three tags (i.e. [v.fut] ~ [v.fut] ~ [v.pres]). Consequently, we create a tag [v.invar] to explicitly mark such instances as undecidable. Identical considerations apply for the respective verbal nouns.

RULE: Replace [v.fut] ~ [v.past] ~ [v.pres] with [v.invar] and replace [n.v.fut] ~ [n.v.past] ~ [n.v.pres] with [n.v.invar].

```
PATTERN:
(\S+\|(?:\[(?!(?:n\.)?v\.)[^\]]*\])*(?:\[(?:n\.)?v\.aux\])?(?:\[(?:n\.)?v\.cop\])?)\[(
n?\.?v\.)fut\](\[\2imp\])?\[\2past\]\[\2pres\](\S*)

REPLACE: $1$3[$2invar]$4
```

(67). The creation of the tags [v.fut.v.past] and [n.v.fut.n.v.past]
BACKGROUND: In many syntactic contexts the rule based tagger will be unable to unambiguously specify the choice of verb stems. For example, *bskon te* could be future or past. Because such contexts are systematically ambiguous it would not be useful to present the human user with a choice between these two tags (i.e. [v.fut] ~ [v.past]). Consequently, we create a tag [v.fut.v.past] to explicitly mark such instances as undecidable. Identical considerations apply for the respective verbal nouns.

RULE: Replace [v.fut] ~ [v.past] with [v.fut.v.past] and replace [n.v.fut] ~ [n.v.past] with [n.v.fut.n.v.past].

```
PATTERN:
(\S+\|(?:\[(?!(?:n\.)?v\.)[^\]]*\])*(?:\[(?:n\.)?v\.aux\])?(?:\[(?:n\.)?v\.cop\])?)\[(
n?\.?v\.)fut\](\[\2imp\])?\[\2past\](\S*)

REPLACE: $1[$2fut.$2past]$3$4
```

(68). The creation of the tags [v.fut.v.pres] and [n.v.fut.n.v.pres]
BACKGROUND: In many syntactic contexts the rule based tagger will be unable to unambiguously specify the choice of verb stems. For example, *mi gśegs* could be future or present. Because such contexts are systematically ambiguous it would not be useful to present the human user with a choice between these two tags (i.e. [v.fut] ~ [v.pres]). Consequently, we create a tag [v.fut.v.pres] to explicitly mark such instances as undecidable. Identical considerations apply for the respective verbal nouns.

RULE: Replace [v.fut] ~ [v.pres] with [v.fut.v.pres] and replace [n.v.fut] ~ [n.v.pres] with [n.v.fut.n.v.pres].

```
PATTERN:
(\S+\|(?:\[(?!(?:n\.)?v\.)[^\]]*\])*(?:\[(?:n\.)?v\.aux\])?(?:\[(?:n\.)?v\.cop\])?)\[(
n?\.?v\.)fut\](\[\2imp\])?\[\2pres\](\S*)

REPLACE: $1[$2fut.$2pres]$3$4
```

(69). The creation of the tags [v.past.v.pres] and [n.v.past.n.v.pres]
BACKGROUND: In many syntactic contexts the rule based tagger will be unable to unambiguously specify the choice of verb stems. For example, *gśegs nas* could be past or present. Because such contexts are systematically ambiguous it would not be useful to present the human user with a choice between these two tags (i.e. [v.past] ~ [v.pres]). Consequently, we create a tag [v.past.v.pres] to explicitly mark such instances as undecidable. Identical considerations apply for the respective verbal nouns.

RULE: Replace [v.past] ~ [v.pres] with [v.past.v.pres] and replace [n.v.past] ~ [n.v.pres] with [n.v.past.n.v.pres].

```
PATTERN: (\S+\|\S*)\[(n?\.?v\.)past\]\[\2pres\](\S*)

REPLACE: $1[$2past.$2pres]$3
```

10.3 Restoring ambiguity when a single form might belong to two distinct verbs

When a single form might belong to two distinct verbs, the rules in section 10.2 efface distinctions which should be preserved. The rules in this section aim to reinstate these distinctions. For example, the second phase will change *źu* [v.fut] ~ [v.past] ~ [v.pres] into *źu* [v.invar], but *źu* [v.fut] [v.pres] belong to the verb 'request' whereas *źu* [v.past] belongs to the verb 'melt'; because the human user will want to be presented with *źu* [v.past] ~ [v.past.v.pres] a specific rule must be created to do this. We make one such rule for each verb stem that has this kind of problem. For the sake the clarify of the presentation, the rules that reambiguate verb stems we place before the rules that reambiguate verbal nouns. Within each section rules are presented according to the part-of-speech-tag which they take as input, in the order [v.invar], [v.fut.v.past], [v.fut.v.pres], [v.past.v.pres].

10.3.1 Verb stem reambiguation rules

(70). [v.invar] > [v.invar] ~ [v.imp] ~ [v.past]

The syllable *phaṅs* can either be the past or imperative of the verb *ḥphaṅ, phaṅs, ḥphaṅ, phaṅs* 'save, economize' or it can be an invariant verb *phaṅs* 'long for, feel loss'.

```
PATTERN: ((?:^|\s)ཕངས་?\|\S*\[v\.invar\])(\S*)
```

```
REPLACE: $1[v.past]$2
```

(71). [v.invar] > [v.fut.v.pres] ~ [v.past]

BACKGROUND: The syllable *źu* is either the present/future of 'ask', or the past of 'to melt'.

RULE: Replace *źu* [v.invar] with *źu* [v.fut.v.pres] ~ [v.past]

```
PATTERN: ((?:^|\s)ཞུ་?\|\S*)\[v\.invar\](\S*)
```

```
REPLACE: $1[v.fut.v.pres][v.past]$2
```

(72). [v.invar] > [v.invar] ~ [v.pres]

72a. BACKGROUND: The syllable *za* is either the present of 'eat' or the invariant verb 'itch'.

72a. RULE: Replace *za* [v.invar] with *za* [v.invar] ~ [v.pres].

72b. BACKGROUND: The syllable *skya* is either the present of 'carry' or the invariant verb 'be gray'.

72b. RULE: Replace *skya* [v.invar] with *skya* [v.invar] ~ [v.pres].

```
PATTERN: ((?:^|\s)(?:ཟ|སྐྱ)་?\|\S*\[v\.invar\])(\S*)
```

```
REPLACE: $1[v.pres]$2
```

(73). [v.invar] > [v.invar] ~ [v.past]

73a. BACKGROUND: The syllable *gśags* is either an invariant verb 'to tighten' or the past of a verb *gśog* 'to split'.

73a. RULE: Replace *gśags* [v.invar] with *gśags* [v.invar] ~ [v.past].

73b. BACKGROUND: The syllable *bor* is either an invariant verb 'to wane, be lost' or the past of a verb *ḥbor* 'discard'.

73b. RULE: Replace *bor* [v.invar] with *bor* [v.invar] ~ [v.past].

73c. BACKGROUND: The syllable *mchis* is either an invariant verb 'to be' or the past of a verb *mchi* 'to go'.

73c. RULE: Replace *mchis* [v.invar] with *mchis* [v.invar] ~ [v.past].

73d. BACKGROUND: The syllable *ches* is either an invariant verb 'believe' (*yid ches*) or the past of a verb *che* 'be large'.

73d. RULE: Replace *ches* [v.invar] with *ches* [v.invar] ~ [v.past].

```
PATTERN: ((?:^|\s)(?:གཤགས|བོར|མཆིས|ཆེས)་?\|\S*\[v\.invar\])(\S*)

REPLACE: $1[v.past]$2
```

(74). [v.invar] > [v.invar] ~ [v.fut.v.pres]
BACKGROUND: The syllable *ḥbyor* is either an invariant verb 'come', or the present (and possibly future) of a verb *ḥbyor* 'adhere'.

RULE: Replace *ḥbyor* [v.invar] with *ḥbyor* [v.invar] ~ [v.fut.v.pres]

```
PATTERN: ((?:^|\s)འབྱོར་?\|\S*)(\[v\.invar\]\S*)

REPLACE: $1[v.fut.v.pres]$2
```

(75). [v.past.v.pres] > [v.past] ~ [v.past.v.pres]
BACKGROUND: The syllable *bźag* is either the past of *ḥjog* 'leave, put aside' or is the present or past of *bźag* 'split, tear' (intrans.) (cf. rule 83).

RULE: Replace *bźag* [v.past.v.pres] with *bźag* [v.past] ~ [v.past.v.pres]

```
PATTERN: ((?:^|\s)བཞག་?\|\S*)(\[v\.past\.v\.pres\]\S*)

REPLACE: $1[v.past]$2
```

(76). [v.past.v.pres] > [v.past.v.pres] ~ [v.past]
76a. BACKGROUND: The syllable *gśags* is either an invariant verb 'to tighten' or the past of a verb *gśog* 'to split'. Because syntactic disambiguation will have already specified some contexts as [v.fut.v.pres] (e.g. *gśags nas*) we must disambiguate [v.past.v.pres] as well as [v.invar], which was handled in rule 73.

76a. RULE: Replace *gśags* [v.past.v.pres] with *gśags* [v.past.v.pres] ~ [v.past].

76b. BACKGROUND: The syllable *bor* is either an invariant verb 'to wane, be lost' or the past of a verb *ḥbor* 'discard'. Because syntactic disambiguation will have already specified some contexts as [v.fut.v.pres] (e.g. *bor nas*) we must disambiguate [v.past.v.pres] as well as [v.invar], which was handled in rule 73.

76b. RULE: Replace *bor* [v.past.v.pres] with *bor* [v.past.v.pres] ~ [v.past].

76c. BACKGROUND: The syllable *mchis* is either an invariant verb 'to be' or the past of a verb *mchi* 'to go'. Because syntactic disambiguation will have already specified some contexts as [v.past.v.pres] (e.g. *mchis nas*) we must disambiguate [v.past.v.pres] as well as [v.invar], which was handled in rule 73.

```
PATTERN: ((?:^|\s)(?:གཞགས|བོར|མཆིས)་?\|\S*)(\[v\.past\.v\.pres\]\S*)
```

```
REPLACE: $1[v.past]$2
```

10.3.2 Verbal noun reambiguation rules

All of the relevant reambiguation rules must be repeated for verbal nouns in addition to verb stems.

(77). [n.v.invar] > [n.v.fut.n.v.pres] ~ [n.v.past]

BACKGROUND: The syllable *źu* is either the present/future of 'ask', or the past of 'to melt'.

RULE: Replace *źu-ba* [n.v.invar] with *źu-ba* [v.fut.v.pres] ~ [v.past]

```
PATTERN: ((?:^|\s)ཞུ་བ་?\|\S*)\[n\.v\.invar\](\S*)
```

```
REPLACE: $1[n.v.fut.n.v.pres][n.v.past]$2
```

(78). [n.v.invar] > [n.v.invar] ~ [n.v.pres]

78a. BACKGROUND: The verbal noun *za-ba* is either the present of the verb 'eat' or the invariant verb 'itch'.

78a. RULE: Replace *za-ba* [n.v.invar] with [n.v.invar] ~ [n.v.pres].

78b. BACKGROUND: The verbal noun *skya-ba* is either the present of the verb 'carry' or the invariant verb 'be gray'.

78b. RULE: Replace *skya-ba* [n.v.invar] with [n.v.invar] ~ [n.v.pres].

```
PATTERN: ((?:^|\s)(?:ཟ་བ|སྐྱ་བ)་?\|\S*\[n\.v\.invar\])(\S*)
```

```
REPLACE: $1[n.v.pres]$2
```

(79). [n.v.invar] > [n.v.invar] ~ [n.v.fut]

BACKGROUND: The syllable *btsog* is either an invariant verb 'to be dirty' or the future of a verb 'smash up'.

RULE: Replace *btsog* [n.v.invar] with *btsog* [n.v.fut] ~ [n.v.invar].

```
PATTERN: ((?:^|\s)བཙོག་པ་?\|\S*)(\[n\.v\.invar\]\S*)
```

```
REPLACE: $1[n.v.fut]$2
```

(80). [n.v.invar] > [v.invar] ~ [n.v.past]
80a. BACKGROUND: According to the dictionaries the syllable *riṅs* is apparently an invariant verb 'to hurry', but is also the past of the verb *ḥdriṅ* 'be distant' seen frequently in the phrase *glo-ba ḥdriṅ* 'be disloyal'.

80a. RULE: Replace *riṅs-pa* [n.v.invar] with *riṅs-pa* [n.v.invar] ~ [n.v.past].

80b. BACKGROUND: The dictionaries agree that there is an invariant verb *gtogs* 'to be included in' and a verb pres. *gtog*, past *gtogs*, fut. *gtog* 'snap'. Thus, the form *gtogs* is itself ambiguous.

80b. RULE: Replace *gtogs-pa* [n.v.invar] with *gtogs-pa* [n.v.invar] ~ [n.v.past].

80c. BACKGROUND: The orthographic form *mchis-pa* is either a nominalized form of the invariant verb *mchis* 'to be', or the past tense of the verb *mchi* 'to go'.

80c. RULE: Replace *mchis-pa* [n.v.invar] with *mchis-pa* [n.v.invar] ~ [n.v.past].

```
PATTERN: ((?:^|\s)(?:རིངས་པ|གཏོགས་པ|མཆིས་པ)་?\|\S*\[n\.v\.invar\])(\S*)

REPLACE: $1[n.v.past]$2
```

(81). [n.v.fut.n.v.pres] > [n.v.fut.n.v.pres] ~ [n.v.pres]
81a. BACKGROUND: The syllable *ḥjug* is the present of the transitive verb 'insert', but is also both present and future of the intransitive verb 'enter'.

81a. RULE: Replace *ḥjug-pa* [n.v.fut.n.v.pres] with *ḥjug-pa* [n.v.fut.n.v.pres] ~ [n.v.pres].

81b. BACKGROUND: The syllable *za* is either the present of 'eat' or the invariant verb 'itch'. Because syntactic disambiguation will have already specified some contexts as [v.fut.v.pres] (e.g. *mi za*) we must disambiguate [v.fut.v.pres] as well as [v.invar], which was handled in rule 78.

81b. RULE: Replace *za-ba* [n.v.fut.n.v.pres] with *za-ba* [n.v.fut.n.v.pres] ~ [n.v.pres].

```
PATTERN: ((?:^|\s)(?:ཟ་བ|འཇུག་པ)་?\|\S*\[n\.v\.fut\.n\.v\.pres\])(\S*)

REPLACE: $1[n.v.pres]$2
```

(82). [n.v.past.n.v.pres] > [n.v.pres] ~ [n.v.past.n.v.pres]
BACKGROUND: The syllable *rtog* is either the present of *rtog, brtags, brtag, rtogs* 'examine' or is an ambiguous present or alternate past of *rtog*(*s*) 'perceive'.

RULE: Replace *rtog-pa* [n.v.past.n.v.pres] with *rtog-pa* [n.v.pres] ~ [n.v.past.n.v.pres]

```
PATTERN: ((?:^|\s)རྟོག་པ་?\|\S*\[n\.v\.past\.n\.v\.pres\])(\S*)

REPLACE: $1[n.v.pres]$2
```

(83). [n.v.past.n.v.pres] > [n.v.past] ~ [n.v.past.n.v.pres]
BACKGROUND: The syllable *bźag* is either the past of *ḥjog* 'leave, put aside' or is the present or past of *bźag* 'split, tear' (intrans.) (cf. 75)

RULE: Replace *bźag-pa* [n.v.past.n.v.pres] with *bźag-pa* [n.v.past] ~ [n.v.pres]

```
PATTERN: ((?:^|\s)བཞག་པ་?\|\S*)\[n\.v\.past\.n\.v\.pres\](\S*)

REPLACE: $1[n.v.past][n.v.pres]$2
```

11 A bit of cleaning up at the end

These are rules that convenient to run after everything else, because they require quite a bit of context.

(84). Precluding *la* as a noun between two imperatives
BACKGROUND: The syllable *la* has many interpretations: the allative case, the allative converb, the stem of the proclausal adverb *lar* 'moreover', and the noun mountain pass. Between two imperative verbs only the allative converb is possible; the noun 'mountain pass' can be precluded in this context (rule 45 already excluded the interpretation of *la* as a case marker in this context).

RULE: If the syllable *la* occurs after one [v.imp] and before another [v.imp] then delete [n.count] from the syllable *la*.

```
PATTERN: (\S+\|\[v\.imp\]\s+ལ་\|\S*)\[n\.count\](\S*\s+\S+\|\[v\.imp\]\s+)

REPLACE: $1$2
```

(85). Finding numbers
BACKGROUND: Earlier rules work to isolate numbers from those morphemes with which they are sometimes homophonous (cf. rule 13), but these rules were written conservatively, requiring an unambiguous number in the immediate context. If three or more morphemes occur in a row, each of which has an analysis as a number, the string of morphemes should together be taken as a number.

RULE: If three or more morphemes occur in a row, each of which has an analysis [num.card], then tag all of them with [num.card].

```
PATTERN:
(\S+)\|\S*\[num\.card\]\S*\s+(\S+)\|\S*\[num\.card\]\S*\s+(\S+)\|\S*\[num\.card\]\S*

REPLACE: $1|[num.card] $2|[num.card] $3|[num.card]
```

12 Evaluation of performance

By removing impossible part-of-speech analyses, the rule-based tagger succeeds in speeding up the process of hand-annotating a new text. In this section, we quantify the gain by evaluating the tagger's performance.

We compare the performance of the rule-based tagger against the baseline performance of our 'lexical tagger'. The lexical tagger draws on a lexicon of words and their possible tags, constructed by combining the verb dictionary with previously tagged text. The lexical tagger consults this lexicon, and assigns to each word all possible analyses of that word. As described above, the output of the lexical tagger serves as input to the rule-based tagger.

The following table compares the performance of the two taggers. We evaluate the taggers against our Classical Tibetan corpus, which currently consists of 76,539 part-of-speech tagged words. The 'Correct' column counts the number of words assigned a single, unambiguous tag, where that tag is also the correct tag for the word. As indicated, the rule-based tagger introduces a 53% gain of 19,058 correct tags over the lexical tagger. Put otherwise, the rule-based tagger correctly tags an additional 25% of the words.

	Tags	*Correct*	*Accuracy*	*Ambiguity*
LexTagger	76,539	36,019	1.000	2.380
RuleTagger	76,539	55,077	0.991	1.395

Following van Halteren (1999), we also evaluate the two taggers along dimensions of 'accuracy' and 'ambiguity'. Accuracy is a measure of the percentage of words that are assigned the correct tag, including those words that are assigned multiple tags where the other tags are incorrect. Ambiguity is a measure of the average number of tags assigned to each word. The aim is of course for both accuracy and ambiguity to be as close as possible to 1.

The table shows that the lexical tagger is 100% accurate. Because every word that occurs in the test data is in the lexicon, with every possible part-of-speech tag for that word listed, each word in a text will be assigned the correct tag (among other, incorrect tags, for lexically ambiguous words). By contrast, the rule-based tagger is 99% accurate. In a small number of cases that we hope to reduce to zero, the rule-based tagger errs by eliminating candidate tags that turn out to be correct.

Turning to ambiguity, while the lexical tagger assigns on average 2.4 tags per word, the rule-based tagger reduces this to 1.4 tags per word.[13] This on average reduction of one tag per word is perhaps more impressive when we factor out those words listed in the lexicon with only one part-of-speech. The total number of tags assigned by the lexical tagger and rule-based tagger are 182,226 and 106,804, respectively. Excluding the 36,019 words with just one part-of-speech, that leaves 146,207 and 70,785 tags for 40,520 words. Computing ambiguity scores for this subset of words gives 3.06 for the lexical tagger and 1.75 for the rule-based tagger.

The rule-based tagger therefore removes an average of 1.31 candidate part-of-speech tags from each lexically ambiguous word. In doing so, the rule-based tagger is able to disambiguate approximately 47% of ambiguous words.[14] The remaining 53% or 21,462 words are assigned 51,727 tags, for an average of 2.4 tags per word. This, too, is a significant improvement from the 3.60 tag per word starting point.

In summary, our conviction that the rule-based tagger substantially improves on the output of the lexical tagger is supported by an evaluation of its performance. While much of the slack will

13 More precisely, the ambiguity scores are 2.3808254615294 for the lexical tagger, and 1.3954193287082 for the rule-based tagger.

14 Computed as 19,058/40,520. The figure is approximate for two reasons: first, it is rounded off; and second, it is possible that a small number of the 36,019 words tagged correctly by the lexical tagger are not included in the rule-based tagger's 55,077 correctly tagged words.

eventually be filled in by a statistical tagger, we anticipate that the rule-based tagger will continue to bring benefit well into the future.

References

Beyer, Stephan V. (1992). *The classical Tibetan language*. Albany: State University of New York Press.

Bsam-gtan (1979). *Dag yig gsar bsgrigs*. Xining: Mtsho-sṅon mi rigs dpe skrun khaṅ.

Derge Kanjur. An electronic edition of the Derge Kanjur prepared by the British Library and the School of Oriental and African Studies. http://www.thlib.org/encyclopedias/literary/canons/kt/catalog.php#cat=d (accessed 23 October 2013)

Garrett, Edward; Hill, Nathan W.; Kilgarriff, Adam; Vadlapudi, Ravikiran; Zadoks, Abel (In Press). "The contribution of corpus linguistics to lexicography and the future of Tibetan dictionaries." *Revue d'Etudes Tibétaines*.

van Halteren, Hans (1999). "Performance of taggers". pp. 81-94 in H. van Halteren (ed.), *Syntactic Wordclass Tagging*. Netherlands: Springer.

Hill, Nathan W. (2010). *A lexicon of Tibetan verb stems as reported by the grammatical tradition*. Munich: Bayerische Akademie der Wissenschaften.

Hill, Nathan W. (2012). "Tibetan -las, -nas, and -bas." *Cahiers de Linguistique Asie Orientale* 41 (1): 3-38.

Schwieger, Peter (2006). *Handbuch zur Grammatik der klassischen tibetischen Schriftsprache*. Halle (Saale): IITBS, International Institute for Tibetan and Buddhist Studies.

Zhang Yisun (1985). *Bod rgya tshig mdzod chen mo*. Peking: Minzu chubanshe.

Edward Garrett
eg15@soas.ac.uk

Nathan W. Hill
nh36@soas.ac.uk

Abel Zadoks
az4@soas.ac.uk