


# The contribution of corpus linguistics to lexicography and the future of Tibetan dictionaries

Edward Garrett\*, Nathan W. Hill\*, Adam Kilgarriff†, Ravikiran Vadlapudi†, Abel Zadoks\*

(\*SOAS, University of London, †Lexical Computing Ltd)

he first alphabetized dictionary of Tibetan appeared in 1829 (cf. Bray 2008) and the intervening 184 years have witnessed the publication of scores of other Tibetan dictionaries (cf. Simon 1964). Hundreds of Tibetan dictionaries are now available; these include bilingual dictionaries, both to and from such languages as English, French, German, Latin, Japanese, etc. and specialized dictionaries focusing on medicine, plants, dialects, archaic terms, neologisms, etc. (cf. Walter 2006, McGrath 2008). However, if one classifies Tibetan dictionaries by the methods of their compilation the accomplishments of Tibetan lexicography are less impressive.

Methodologies of dictionary compilation divide heuristically into three types. First, some dictionaries lack explicit methodology; these works assemble words in an *ad hoc* manner and illustrate them with invented examples. Second, there are dictionaries that are compiled over very long periods of time on the basis of collections of slips recording attestations of words as used in context. Third, more recent dictionaries are compiled on the basis of electronic text corpora, which are processed computationally to aid in the precision, consistency and speed of dictionary compilation. These methods may be called respectively the 'informal method', the 'traditional method', and the 'modern method'. The overwhelming majority of Tibetan dictionaries were compiled with the informal method. Only five Tibetan dictionaries use the traditional methodology. No Tibetan dictionary yet compiled makes use of the modern method.

## The Informal Method

The vast majority of Tibetan dictionaries in no way specify the methods by which they were compiled; the principle for excluding or including words and the evidence for the correctness of definitions remain tacit. The failure of the compilers of such dictionaries to have considered the need for principled decision-making probably

accounts for this silence. When writing a dictionary is approached pre-theoretically, the failings of existing works normally serve as the impetus for a new work's compilation. However, lacking a robust procedure by which to achieve progress the would-be lexicographer reenacts the errors of his forebears. Dictionaries of this type, no matter their girth, amount to glossaries or hand-lists; they are informal affairs prepared as aides to reading and not as works of scholarship. Specifically, such projects fail on three grounds: 1. They rely on previous dictionaries. 2. They rely on intuition rather than evidence. 3. They use invented examples rather than genuine examples.

Because it is the most recent large scale Tibetan-English dictionary, which many turn to as a first port of call, Melvyn Goldstein's (2001) *A New Tibetan English Dictionary of Modern Tibetan* serves as a convenient example of the informal method. A perceived insufficiency in the coverage of modern and administrative terminology and insufficient overall number of entries inspire Goldstein's project (2001: vii). Without an explicit methodology Goldstein is however unable to implement his concentration on the modern language. The work includes Old Tibetan words such as *hon-te* 'but, however', *hu-bu-cag* 'we', and *liñs* 'hunt' without comment; without acknowledgment these words are included on the testimony of previous dictionaries.

A dictionary that merely reports the information contained in other dictionaries in principle spares the user the need to consult those dictionaries on which it is based. The explicit goal of Hill (2010) is to present the testimony of previous dictionaries in order to spare the user the time of looking up the same word in nine earlier sources. However, whereas Hill (2010) specifies which morphological forms and which meanings are found in which previous sources, Goldstein omits such information. By not distinguishing entries based on the compilation of primary sources from entries incorporated wholesale from other works Goldstein obfuscates the value of the former and disguises the latter as his own contribution.

Impressionistically, Goldstein relies above all on the *Bod rgya tshig mdzod chen mo* (Zhang 1985). For example, if one compares the treatment of words starting with *liñ-*, Goldstein adds two Chinese loanwords and subtracts a few words to do with Buddhism, but the reliance on Zhang (1985) is otherwise evident in every entry. Unfortunately, in Goldstein's hands a *réchauffage* of Zhang's entries loses value. For example, whereas the *Bod rgya tshig mdzod chen mo* signals that *hon-te* is archaic (*rñiñ*), Goldstein inexplicably excludes this stipulation. In short, the savings of time achieved by consulting Goldstein rather than his sources entails a concomitant loss.

The criticisms presented here have focussed on Goldstein but could be readily applied to nearly any other Tibetan dictionary. The *Bod rgya tshig mdzod chen mo* itself also fails to specify sources or methods, and does not cite real examples. Although works compiled with an informal method are of undeniable practical benefit as a place to look for a word one comes across in a text, their pragmatic benefit (and not only their scholarly value) is compromised by their methodological failings. A student looking for a common word must flip past rare or dubious words. A researcher looking for the meaning of a technical term is confronted by many homonyms listed together with no specification of genre or time period, even though no Tibetan text would use both. If the scope of a dictionary were explicitly stated and tightly controlled, such time wasting obstacles would be avoided. Thus, although the description of Goldstein's as "one more in a long parade of ignorant, mistake-filled books on the Tibetan language" (Beckwith 2001: 398) may appear a harsh verdict for a volume that many students find beneficial, as a call for future dictionaries to aim higher Beckwith's verdict is well founded.

#### The Traditional Method

Excerpting wholesale from previous lexicographical works commits the editor of a new dictionary to all of the mistakes of his predecessors. To avoid this pitfall, previous works should be seen as providing hypotheses, but these hypotheses must be tested against a body of data. In the traditional approach to lexicography a team of readers reads through a set of texts and writes onto slips of paper attestations of a word in its context together with a citation sufficiently explicit to find the passage again. The slips thus created are then filed according to alphabetical order in boxes or cabinets until a sufficient number of slips is available to provide a good set of data for establishing the meanings of the words. At this point in the process the team writes up the dictionary entries availing themselves of the collection of slips.

This method has led to many great dictionaries which are monuments of human achievement, but the traditional method is very slow. The Grimm brothers *Deutsches Wörterbuch* began in 1838 and reached completion in 1961. The *Oxford English Dictionary* started in 1857 and the complete first edition was brought out in 1928. The *Thesaurus Linguae Latinae* began in 1894 and published volume P in 2010. The *Wörterbuch der ägyptischen Sprache* began in 1897 and finished in 1963. The Chicago Assyrian Dictionary began in 1921 and reached completion in 2011. The slow pace of work that the traditional method requires, inevitably leads to many dictionaries being abandoned after a few letters. The Burmese-English dictionary

begun by J. A. Stewart in 1925 was abandoned after one letter in 1981. The Pennsylvania Sumerian dictionary begun in 1974 abandoned publication in 1994 after the letter B. Beginning at the beginning of the alphabet and moving painstakingly through alphabetical order also has the disadvantage that editorial decisions taken lightly in the beginning can hamstring the project for decades to come.

Despite the high level of results that the traditional method achieves, it still suffers methodologically from what one might want to achieve. In general a reader preparing slips will be drawn to contextually salient words and thereby the evidence for very common words may be thin. To counteract this tendency some works may be fully indexed to ensure sufficient coverage of common words. Nonetheless, thousands of possible attestations of common words must simply be ignored for lack of space and time, without in any way informing the analysis of these words.

Because readers will by definition encounter rare words only rarely, it is also difficult to collect sufficient slips in such cases. To militate against this obstacle one may have readers focus their attention on particular genres in which words of overall rarity will be comparatively more common. Nonetheless, this is a half measure. There is no way in the traditional method for a reader to narrow his focus onto rare words *per se*.

Perhaps more troubling from a methodological perspective, in the traditional method there is no way of even knowing what the true frequency of a word is, because no record is made of the vast majority of words seen by the readers. One may hope that the number of slips collected well reflects the frequency and behavior of a particular word in the works consulted, but there is no way to know if this hope is realized.

Five Tibetan dictionaries can be said to conform to the traditional approach. First, Jäschke (1881) provides clear citations of original texts in support of his definitions. It is always clear what citation supports which claim and the strength with which a claim can be made is also made explicit. Jäschke (1881) is a lexicographical work of the highest standard and is still profitably consulted today. Second, Lokesh Chandra compiled a 12 volume Tibetan-Sanskrit dictionary on the basis of canonical Buddhist texts available in both languages (1958-61). This work was continued with seven supplementary volumes (1992-1994) and a one volume Sanskrit-Tibetan index (2007). Third, Negi (1993-2004) compiled another Tibetan-Sanskrit dictionary, this one in sixteen volumes. Negi includes extensive quotations in addition to citations and made reference to a larger

number of texts than Chandra.<sup>1</sup> Fourth, Nag dbaṅ tshul khrims (1997) provides a dictionary of difficult or archaic words. He provides attestations and cites the works they are found in, but does not specify page and line numbers and has an inadequate bibliography; consequently, these citations are not easily verified. Fifth, the single most impressive work of Tibetan lexicography is the ongoing *Wörterbuch der tibetischen schriftsprache* published by the Bayerische Akademie der Wissenschaften (Francke et al. 2005-). Helmut Hoffmann founded the project in 1954; the first fascicle was published in 2005. The sixteen fascicles published by 2011 cover from *ka* until *gcags*. Each entry gives copious citations of original sources precisely cited to page and line number. The use of previous dictionaries is carefully distinguished from the evidence of textual attestations. In addition, very thorough reference to previous scholarship is given when relevant.<sup>2</sup>

#### The modern method

The availability of electronic text editions greatly facilitates the collation of the lexical attestations that form the bedrock of the traditional method. Once a text is available in an electronic version the need for a human reader to meticulously read through the text, copying out attestations with their contexts onto paper slips, disappears. Instead, with the click of a button a researcher can assemble all attestations available within that text in context. Several of the drawbacks of the traditional have ceased to exist. Since the phase of slip collection can essentially be skipped, the process of compiling a dictionary becomes much faster.

There is much more to the modern method than the speeding up of collecting attestations through the availability of e-texts. These e-texts themselves introduce problems of their own; the modern method includes the use of e-texts and the solutions to those problems that the use of e-texts introduces.

Using the traditional method the lemma list that will serve as headwords in the dictionary is compiled in an ongoing way as

- 
- 1 In addition to these two Tibetan-Sanskrit dictionaries, there are bilingual indices available for a number of Tibetan translations of Sanskrit Buddhist texts: *Abhidharmakośabhaṣya* (Hirakawa 1973-1978), *Kāśyapaparivarta* (Weller 1933), *Mahāyānasamgraha* (Nagao 1994), *Mahāyānasūtralaṅkāra* (Nagao 1958-1961), *Meghadūta* (Chimpa et al. 2011), *Nyāyabindu* (Obermiller 1970), *Prasannapadā Mādhyamakavṛtti* (Yamaguchi 1974), *Yogācārabhūmi* (Yokoyama 1996), *Laṅkāvatārasūtra* (Suzuki 2000), *Sukhāvativyūha sūtra* (1984), *Saddharmapūṇḍarikasūtra* (Ejima et al. 1985-1993), among others.
  - 2 The compilation of the dictionary is discussed by Uebach & Panglung (1998), to which Maurer & Schneider (2007) and Schneider & Maurer (2012) provided a more recent perspective.

readers file slips in the cabinet. But to search in an electronic corpus one must already know what to search for. That is, there has to be a predefined lemma list. However, using a predefined lemma list precludes the discovery of new vocabulary items. Instead, we must find a way to have the e-texts themselves tell us the lemma list.

Another problem which e-texts introduce is the availability of too much data. In the traditional method, the readers pre-select what goes onto the slips, but searching an electronic corpus will yield all examples that meet the search parameters. Depending on the size of the electronic corpus this is more or less of a problem. Increasing the size of the corpus exacerbates the problem, but even with a small corpus the attestations of common words are overwhelming in number. Consideration of two available online corpora of Tibetan texts will illustrate these principles. The Old Tibetan Documents Online (OTDO) is a collection of 109 Old Tibetan texts.<sup>3</sup> The eKanjur is an electronic version of the Derge Kanjur, that unfortunately contains many typos.<sup>4</sup>

A lexicographer investigating the behavior of the indefinite article with the three sandhi forms *cig*, *zig*, and *sig* in the OTDO will find 221 examples of *cig*, 480 examples of *zig*, and 330 examples of *sig* (accessed 15, November 2012). The OTDO website provides no method to search for all three forms at once, nor is a way provided to weed out examples of the imperative verb final marker that happens to have the same three forms. This search does not have good precision (the fraction of retrieved instances that are relevant) because it includes many examples of the imperative verb final marker. This search also does not have good recall (the fraction of relevant instances that are retrieved) because a search for *cig* will not retrieve examples of *zig* and *sig*, nor will a search for *zig* yield *cig* or *sig*, etc. Thus, to write a dictionary entry for the indefinite article having looked at all of the relevant evidence in the OTDO one would have to read through a total of 1031 examples - possible, but not convenient. If we turn to the eKanjur the situation gets much worse. There are 14,801 examples of *cig*, 14,011 examples of *zig*, and 7,354 examples of *sig*, making a combined sum of 36,166, far more than any person could possibly look through (accessed 15, November 2012). The need to automatically differentiate the indefinite article from the imperative marker is demonstrable and the ability to search for *cig*, *zig*, and *sig* in one go would be helpful.

The Tibetan spelling *mi* (*myi* in Old Tibetan) signifies two words; one is the noun 'person' and the other a marker of negation. In the

---

3 <http://otdo.aa.tufs.ac.jp/>

4 <http://www.thlib.org/encyclopedias/literary/canons/kt/catalog.php#cat=d/k>

OTDO the spelling *myi* occurs 1,718 times, and in the eKanjur the spelling *mi* occurs 35,434 times. In either corpus a lexicographer looking for examples of 'person' would have to scroll through many screens of 'not'; the search has terrible precision. Although technically the search provides perfect recall, the lexicographer could never be sure to have looked at all examples within a reasonable amount of time.

Negation is common enough in Tibetan texts that a lexicographer interested in *mi* 'not' may not find the occasional example of *mi* 'person' inconvenient. Nonetheless, he may find the overwhelming number of *mi* 'not' inconvenient. It would be sensible in a dictionary entry on *mi* 'not', to provide separate examples of the marker before presents, futures, imperatives (in the *potentialis* function, cf. Zeisler 2002), and adjectives. There is no way to differentiate these uses of *mi* with either the OTDO or the eKanjur. When working with paper slips in cabinets the constraining factor is the labour it takes to assemble examples; with electronic texts the constraining factor is the labour it takes to look through the surfeit of available examples. Some of the burden of classifying and analyzing the examples must be passed from the human lexicographer to the computer.

A computer can be taught to distinguish one word from another (tokenizing), can learn to assign a part-of-speech category to each word (POS-tagging), and learn to associate different orthographic or grammatical forms of the same word (lemmatization). There has been more than twenty years of work on these tasks of tokenizing, POS-tagging, and lemmatization for languages such as English, but for Tibetan such research is still at an early stage. After tokenization, POS-tagging, and lemmatization the frequency of different words are immediately available for calculation. The dictionary project can pick *a priori* whether it will focus on frequent or rare words and determine frequency thresholds for the inclusion or exclusion of vocabulary. If the texts making up the corpus are labelled for genre it is also possible to know with certainty which words occur more frequently in which genres, and the words may be labeled accordingly.

<b>part-of-speech</b>	<b>Number of Examples</b>
Singular noun	7150
Plural noun	1861
Proper noun	4
Past participle form of verbs	572
Past tense form of verbs	139
Infinitive form of verbs	132
Finite base form of verbs	107
The -ing form of verbs	85
The -s form of verbs	51
Unclassifiable	2

Table 1: part-of-speech categories for 'chair'

A look at what can now be easily done with English will give an impression of what may be possible one day in Tibetan. The English examples will make use of the British National Corpus, a part-of-speech tagged corpus of 96,048,950 words, and the Sketch Engine querying software.<sup>5</sup> An English lexicographer will want to distinguish 'chair' (noun) the piece of furniture and 'chair' (verb) a meeting. For the noun he will want to find examples of the plural 'chairs' and well as the singular 'chair' and for the verb he will want to find examples of forms such as 'chaired', 'chairing', and 'chairs'. Instead of checking for each of these separately and reading through ambiguous forms by hand one can simply stipulate the part-of-speech in the search window. More fine-grained categories such as 'past participle' can also be specified (cf. Table 1).

The lexicographer may still find himself with too many examples. A total of 9011 examples of 'chair' as a noun is still a lot to read through. One could of course simply choose one or two that seem easily excerptable, but it would be preferable to know that the examples chosen for the dictionary are somehow typical. The Sketch Engine software solves this problem by creating a 'sketch' of a word's behavior (cf. Table 2). Remaining with 'chair' as a noun, the Sketch Engine tells us what people most often do to chairs (chair as the object of verbs), what chairs themselves do (chair as the subject of

---

5 <http://www.sketchengine.co.uk/>



object_of	Freq	Score	subject_of	Freq	Score	modifier	Freq	Score
swivel	14	8.07	creak	7	8.5	rocking	45	8.83
rock	14	7.43	face	23	4.92	high-backed	37	8.6
upholster	7	7.26	surround	6	4.27	swivel	36	8.53
push	66	7.1	stand	15	3.97	wicker	35	8.42

Table 2: A simplified word sketch of 'chair' (noun)

Lemma	Freq	Score
seat	10462	0.26
bed	16797	0.22
table	22162	0.21
furniture	3457	0.19
armchair	900	0.18

Table 3: A simplified thesaurus entry for 'chair' (noun)

verbs), and what kind of chairs there are (adjectives that modify 'chair').

The Sketch Engine can also find words that have a similar statistical behavior to 'chair'. Such words are normally close in meaning to the word in question. This is a way of developing a thesaurus, but rather than relying on a subjective notion of how close the meaning of two words is, it furnishes a rigorous comparison of two words' statistical behavior (cf. Table 3). Much more can also be achieved using a part-of-speech tagged corpus. Software can compare the word sketch of two different words, suggest examples for inclusion in the dictionary, compare usage across genres or time.<sup>6</sup> Outside of dictionary compilation, part-of-speech tagged corpora are a *sine qua non* for many language related technologies such as automatic translation, speech recognition, auto-completion, optical character recognition, etc. If these technologies are ever to be available to Tibetan speakers then more must be done to create part-of-speech tagged Tibetan corpora.

#### Tibetan in Digital Communication

Tibetan in Digital Communication is a research project funded by the Arts and Humanities Research Council and based at SOAS,

6 For the 'Sketch Engine' see Kilgariff et al. (2004); for the use of corpus tools in lexicography see Kilgariff and Kosem (2012).

University of London, engaged in building a 1,000,000 syllable part-of-speech tagged corpus of Tibetan texts spanning the language's entire history. In addition to the corpus, the project is developing a number of digital tools allowing the corpus to be employed in many areas of humanities research, and enabling other researchers to more easily develop their own corpora or software tools. Tokenization and part-of-speech (POS) tagging fall within the goals of the project, but we do not currently intend to work on lemmatization.

The corpus will itself be a powerful resource for scholars working with Tibetan language materials in a wide range of disciplines—including history, religion, literature and linguistics—since it offers ready access to, and comparison across, texts from different time periods, regions and genres. It will also provide an important foundation for subsequent work on a historically comprehensive, lexicographically rigorous dictionary of Tibetan.

In the following sections, we describe the results of the project so far. On the one hand, we have developed a categorization of the parts of speech for Classical Tibetan based on a pilot study of the first 17,522 words of the *Mdzañs blun*. On the other, we have created a web-based software infrastructure for Tibetan natural language processing focused on tokenization, part-of-speech tagging, and the comparative evaluation of multiple models.

#### POS-Tagging

Our project has categorized the parts of speech of Classical Tibetan based on the first 17,522 words of the *Mdzañs blun*. The categorization presented here is not a rigorous analysis of Tibetan part-of-speech categories made on linguistic principles, instead it is a set of pragmatic solutions to the problems that arose during the tagging of the *Mdzañs blun* pilot corpus. In some cases (e.g. the analysis of *phrag* as a cardinal number), no researcher would agree intellectually with the decision we have made. Nonetheless, we hope that this categorization will be useful, whether for practical implementation in corpus research or as a stepping stone toward a more rigorous linguistic analysis.

For the duration of our project this set of POS-tags will remain fixed so far as practicable and be used in the tagging of the remainder of the Classical Tibetan portion of our corpus. The Old Tibetan and Modern Tibetan components of the corpus will use their own POS-tag sets, but these will also be developed with the tag-set described here as a point of departure. In this presentation, we discuss the POS categories according to the broad syntactic headings 'nouns', 'pronouns', 'adjectives', 'numerals', 'trailing members of the noun phrase', 'adverbs', 'negation', 'verbs', 'affixes', and 'clitics'.

## Nouns

We distinguish four types of nouns: lexical nouns, proper nouns, relator nouns, and mass nouns. We also distinguish verbal nouns from verb stems; verbal nouns are discussed further below together with verbs. Since any verbal noun can be used as a noun, it can be difficult to distinguish between nouns and verbal nouns. When a morphological verbal noun refers to a real physical thing in the world, *hgro-ba* 'animal', *skyes-pa* 'person', *mkhas-pa* 'wise man', *rgyal-ba* 'victor', *bzah-ba* 'food', we tag it as a lexical noun. When it refers to an abstract notion *bsam-pa* 'thought', *bde-ba* 'happiness', etc. we tag it as a verbal noun.

## Lexical nouns (n)

To identify lexical nouns we rely on the syntactic ability of the word in question to head a noun phrase, the dictionary meaning, and (when possible) the presence of nominal suffixes such as *-mo*, *-po* and *-bu*. Because we treat grammatical affixes as separate words, a single word normally does not include grammatical affixes such as case markers and converbs. Nonetheless, there are well-motivated exceptions to this policy like *gañ-na-ba* 'whereabouts' and *bdag-gi-ba* 'that which is mine'.<sup>7</sup> Another frequent category of exceptions is calques of Sanskrit terms, e.g. *kun-tu-rgyu* 'parivrājaka', *rten-ciñ-hbrel-bar-hbyuñ-ba* 'pratīyasamutpāda', etc.

In general, when two nouns occur in succession they are understood as a compound; the *dandva* compound *pha-ma* 'parents' is treated as one noun rather than two (*pha* 'father' and *ma* 'mother') and the *tatpuruṣa* compound *khyim-bdag* 'householder' is likewise treated as one noun rather than two (*khyim* 'home' and *bdag* 'lord'). When an adjective precedes its head this is also treated as a compound. Thus, because *dug btsan-po* would be the expected order for 'mighty poison', we treat *btsan-dug* 'mighty poison' (vol. 74, page 147a) as a single word.

Apposition is the one category of exceptions when the concatenation of two nouns is not treated as a compound. An example of this type is the two words *bu khyeḥu* in the following sentence: *deḥi tshe yul de na khyim-bdag cig la bu khyeḥu žig btsas na /* 'At that time, in that land, when a child, a son, was born to a householder,' (vol. 74, page 142b). Rather than understanding *khyeḥu* as an adjective modifying *bu*, or taking *bu-khyeḥu* as one word, the second word simply sits after the first one to add greater specificity.

7 We treat a *-pa* or *-ba* as part of the preceding word, regardless of the part-of-speech of the preceding word.

The reason to not treat apposition as compounding is because apposition occurs with proper nouns. For example, in the sentence *dehi tshe na rgyal-po Gsal-rgyal gyi btsun-mo chen-po Hbar-li zes bya-ba la bu-mo zig btsas-nas* 'At that time a daughter was born to Hbar-li, the main queen of king Prasenajit' (vol. 74, page 148a), *rgyal-po* 'king' and *Gsal-rgyal* 'Prasenajit' are in apposition; to unite the two as a compound would lead to unintuitive, unwieldy, and therefore unacceptable consequences.

Proper nouns (n.prop)

This tag is used for personal and place names.

Relator nouns (n.rel)

A relator noun is a noun, normally one syllable, which has a genitive before it and a spatial case (allative, locative, terminative) after it, e.g. *dehi nañ na* 'inside of that', *dehi druñ du* 'before him', *dehi hog tu* 'under that', *dehi tshe na* 'at that time'; relator nouns are not quantified and are not suffixed with adjectives, determiners or demonstratives. After identifying the class of relator nouns, these words are tagged as relator nouns even in syntactic contexts missing the genitive to left or the spatial case to the right. For example, in the sentence *dehi tshe blon-po zig phyi-rol nas nañ du hoñs-pa las/ mi btson du bzuñ-ba mthoñ-ba dan /* 'Then the minister went inside from outside and saw the man who had been taken to prison' (vol. 74, page 147a), the relator noun *tshe* is not followed by a spatial case and the relator noun *nañ* is not preceded by a genitive.

In the phrase *bar hgañ* 'sometimes' we do not consider *bar* a relator noun because it undergoes quantification, e.g. ... *bar hgañ ni gti-mug gi phyir lus btañ yañ chos kyi phyir bsod-nams kyi zin dan lan hgañ yañ ma phrad-pañ lus hdi ci ruñ ?* 'What is the use of this body which ... sometimes has been used because of ignorance, but has not yet met an occasion (to serve) as a field of merit' (vol. 74, page 139b).

Mass nouns (n.mass)

We divide out mass nouns from normal lexical nouns on the basis of two instances in our corpus where otherwise two nouns not in apposition would follow each other: *nor-bu sbar gañ* 'a handful of jewels' (vol. 74, page 153b) and *chu sñim-pa gañ* 'a handful of water' (vol. 74, page 144b). Knowing that there exists this syntactic difference between normal lexical nouns and mass nouns, we tag all plausible mass nouns on the basis of their meaning (e.g. *zañs* 'copper'). A final list of mass nouns can only be securely put forward after the syntactic behavior of these words is better investigated.

## Pronouns

We distinguish three types of pronouns: indefinite, interrogative, and personal.

## Indefinite pronouns (p.indef)

The words *la-la* 'some', *so-so* 'each', and *gñi-ga* 'both' are used as indefinite pronouns in our pilot corpus.

## Interrogatives (p.interrog)

This is the tag used for words such as *su* 'who', *nam* 'when', and *gañ* 'where'.

## Personal pronouns (p.pers)

The first 17,522 words of the *Mdzans blun* include the first person pronouns *na*, *bdag-cag*, and *kho-bo* and the second person pronouns *khyod* and *khyed*. The tagset does not distinguish the person and number of personal pronouns.<sup>8</sup>

## Adjectives (adj)

It is difficult to distinguish a class of words in Tibetan which are unambiguously adjectives. Those words that can occur immediately after a noun may appear morphologically to be nouns (*chen-po* 'big', *bzai-po* 'good', etc.) or verbal nouns (*che-ba* 'big', *mdzes-pa* 'beautiful'). Because all verbal nouns can function adjectivally but not all adjectives are verbal nouns we draw a distinction between adjectives properly speaking (those that are not verbal nouns) and verbal nouns functioning as adjectives. The latter, regardless of their frequency in attributive position, are tagged as verbal nouns. The category of adjectives is thus defined negatively vis-à-vis verbal nouns, adjectives are a morphologically heterogeneous class.<sup>9</sup>

Distinguishing between adjectives and verbal nouns is not always easy. When the removal of the *-pa* suffix yields a verb stem, there is no objection to calling the form with the *-pa* a verbal noun. However, in the case of the adjectives *g.yas-pa* 'right' and *gcig-pa* 'alone' the removal of the *-pa* leaves *g.yas* and *gcig* which are not verbal stems. Consequently, we tag *g.yas-pa* and *gcig-pa* as adjectives and not as verbal nouns.

---

<sup>8</sup> About personal pronouns in Old and Classical Tibetan see Hill (2007, 2010).

<sup>9</sup> Adjectives can be used as nouns directly, what one might analyze as omission of the head noun, in order to avoid proliferating each adjective into a noun and an adjective we tag this as adjective as well.

However, if the removal of *-pa* leads to a word that is not known to us as a verb, but is also not unambiguously non-verbal, we continue to regard the word as a verbal noun. In other words, when in doubt, the annotator favors the interpretation of a word as a verbal noun, but in the cases *g.yas-pa* and *gcig-pa* there is no doubt. This circumstance has led to the following words being tagged as verbal noun in the hand tagged portion of the *Mdzañs blun*.

- phañs-pa* 'dear, beloved' (no known verbal equivalent)
- khol-pa* 'boiling' (distinct from all tenses of the verb 'khol 'boil')
- bzad-pa* 'tolerable' (occurs in *mi bzad-pa* 'intolerable' which suggests it is a verb 'to tolerate'. The dictionaries however give *bzod* 'tolerate' with no verb *bzad*).
- skems-pa* 'lean' (distinct from all tenses of the verb *skem* 'dry')
- skam-pa* 'dry' (distinct from all tenses of the verb *skem* 'dry')
- bśor-ba* 'mangy' (distinct from the verb *bśor* 'hunt')

We expect that in some cases consideration of further data will vindicate the analysis of these words as deriving from verbal stems.

Although words such as *nag* 'black' and *gsar* 'new' are frequently treated like adjectives for pedagogical purposes, a single syllable in predicate position before verbal suffixes is a verb. These words may appear to occur attributively, but we see this as the formation of compounds. The compound *blon-chen* 'prime minister' contrasts beautifully with noun and adjective pair *blon-po chen-po* 'great minister'. In this case the *-n* in the word shows that the *chen* in *blon-chen* is not the verbal stem, but rather a short form of the adjective. This clarity is lost in a compound like *thig-nag* 'black dot' which is a compound version of *thig-le nag-po* and not a use of the verb *nag* 'be black'. The same distinction can be drawn between the compound *bu-chuñ* 'small child' versus *bu chuñ-ñu* or *bu chuñ-chuñ*.

#### Numerals

In numbers we distinguish cardinals (*gcig*, *gñis*, *gsum*, etc.) and ordinals (*dañ-po*, *gñis-pa*, *gsum-pa*, etc.). Other derivatives of numerals are treated according to their respective syntax, thus *gcig-pa* 'sole' is an adjective, *gñi-ga* 'both' is an indefinite pronoun, etc. In higher numbers each component digit is tagged separately, to do otherwise would prevent the computer from learning patterns by virtue of having to independently learn each possible cardinal number of the infinite possibility.

When a numeral follows a noun we regard the two as separate words. In addition to obvious cases like *mi lña* 'five men', we also treat *dkon-mchog gsum* 'triratna' as two words. While it is true that one

will almost never encounter any other numeral after the word *dkon-mchog* this fact says as much about Buddhism as it does about syntax.<sup>10</sup>

The treatment of *phrag* well exemplifies our pragmatic attitude toward part-of-speech tagging. Although not a cardinal number itself, this syllable occurs inside cardinal numbers, effectively marking a certain place with a zero, e.g. *stoñ phrag drug cu* '1060'. Because the internal structure of numerals is not of interest to our project and adding a new tag for *phrag* would add unnecessary complications to our tag-set, we treat *phrag* itself as a cardinal number.

Trailing members of the noun phrase

In a conceptually imprecise category, marked with the letter 'd' at the beginning of a POS-tag, we group together those classes of word that occur in the noun phrase after nouns and adjectives, but before case markers. The choice of the letter 'd' is arbitrary, but invokes the fact that demonstratives and determiners are members of this category. The subdivisions of this group are demonstratives (d.dem), determiners (d.det), emphatics (d.emph), the indefinite (d.indef), and plurals and quantifiers (d.plural).

Demonstratives (d.dem)

This tag is used for the demonstratives *h̄di* 'this' and *de* 'that'. These two words are tagged as demonstratives also when used as determiners (i.e. we do not distinguish *rgyal-po de* 'that king' from *de* 'that one, him').

Determiners (d.det)

The most frequent determiner is *ḡzan* 'other'. In addition, we identify *ya-re* 'each one (of two)' as a determiner on the basis of the following sentence: *Brgya-byin dañ Tshais-paḥi rgyal-pos lag-pa ya-re nas zin te* 'The kings Indra and Brahma each took him by one of his hands' (vol. 74, page 135a). We reckon *h̄baḥ* 'sole' as a determiner on the basis of sentences such as *rus-pa dañ khrag h̄baḥ žig gis sa rtsog-rtsoḡ ltar h̄dug-pa mthoñ* 'They saw the ground besmirched with only bone and blood' (vol. 74, page 139b).

---

10 There are occasions when the morphology of a word suggests that it might contain a numeral (e.g. *m̄non-sum* 'real', *phun-sum-tshogs* 'marvelous'), but there is no reason to see such cases as synchronically analyzable.

## Emphatics (d.emph)

We initially invented this category for *ñid* in phrases such as *rgyal-po ñid* 'that very king' or *lus ñid* 'this body'. This syntactic use of *ñid* must be distinguished from its use in Buddhist terminology -*ñid* inside of words, e.g. *stoñ-pa-ñid* 'emptiness'. Apart from *ñid*, we have categorized *kho-na* 'the very, same' and *re-re* 'each' as emphatics.

This use of *kho-na* should not be confused with its function as a third person pronoun in Old Tibetan. In one case *kho* appears not as a personal pronoun but as what seems to be a variant of *kho-na*; this *kho* we also classify as an emphatic, viz. *smras-paḥi tshig ḥdi bden na bden-paḥi tshig bden-paḥi tshig smras-pas / bdag gi lus ḥdi sñā-ma kho bžin du rma med-par gyur cig* 'If these words that I have said are true, then because of saying true words, let this my body be without wounds like before' (vol 74, page 137b).

## Indefinites (d.indef)

This category is used for the allomorphs of the indefinite marker *cig*, *žig*, and *śig* as in *pho-ñā cig* 'a messenger'. The indefinite marker, which occurs inside of noun phrases, must be distinguished from the identically looking imperative converb (see below), which occurs suffixed to the imperative stems of verbs.

## Plurals and Quantifiers (d.plural)

The plural markers *rnam*s, *dag*, *kun*, *thams-cad*, *ḥo-cog* (and its variants) and *tsho* are tagged as their own category 'plural'. However, plural pronouns (*bdag-cag*, *khyed-cag*, *ḥu-bu-cag*) are treated as one word. The plural marker *-cag* is not removed because to do so would result in pronominal stems which are not mutually comparable (viz. *bdag* is a singular pronoun, *khyed* a plural pronoun, and *ḥu-bu* has no independent life outside of *ḥu-bu-cag*). We also tag *ḥgah* 'some' as a plural, although in the abstract one would perhaps prefer to call it a 'quantifier'.

The three verbs (*la*) *sogs-pa* 'etc', (*dañ*) *ldan-pa* 'having', (*dañ*) *bcas-pa* 'together with' could be seen as similar to quantifiers or otherwise to be treated as parts of the noun phrase, however, we have chosen to treat them etymologically as verbs.

## Adverbs

We distinguish four types of adverbs: 'directional' (adv.dir), 'temporal' (adv.temp), 'intensive' (adv.intense), and 'proclausal' (adv.proclausal).



## Directional adverbs (adv.dir)

We use this tag for adverbs that end in *-cad*, i.e. *phyin-cad* 'after', *sñon-cad* 'before', *man-cad* 'below', *yan-cad* 'above', *slan-cad* 'after'. We also include *phan-tshun* 'mutually' in this category, for lack of a better place to put it.

## Temporal adverbs (adv.temp)

Temporal adverbs are those that occur in syntactic positions or have morphological structure that suggests they are nouns, that refer to time, and that are not followed by case markers. In our corpus so far the temporal adverbs are *sñon* 'previously', *da* 'now', *deñ* 'these days', *mdañ* 'yesterday', *gdod* 'at first', *da-ruñ* 'still', *phyi-ñin* 'the next day', *phyi-dro* 'in the afternoon', and *sañ* 'the next day'. There are also nouns that refer to time such as *žag* 'day' and *gdugs* 'noon', but these behave syntactically as nouns, for example by being suffixed with case markers. In the phrase *sañ gi gdugs la*, where *sañ* also appears to function as a noun, we have hesitatingly decided that for the time being it is best to tag *sañ* as a temporal adverb.

## Intensive adverbs (adv.intense)

This tag is used for the adverbs *rab (tu)* 'very' and *śin (tu)* 'very', because as uninflected stems they do not occur independently.

## Proclausal adverbs (adv.proclausal)

A small number of words occur clause initially and refer to the content of the previous clause. Such adverbs often begin with a demonstrative stem. These words are classed as 'proclausal adverbs'. In the following list they are presented together with their affixal suffixes, but in our tagging we divide off these suffixes in order to be consistent with their treatment elsewhere: *de (nas)* 'then', *de (ste)* 'thereafter', *gal (te)* 'if', *ho (na)* 'in that case', *hon (te)* 'nevertheless', *yañ (na)* 'alternatively'.

## Negation (neg)

The two negation prefixes *ma* and *mi* are classified together in their own category. In the modern language and presumably in its ancestors these morphemes combine phonologically with the following word. However, treating them as separate words has the advantage of reducing the number of tags and simplifying the task of word-breaking. For the two verbs *min* and *med* negation is inherent to their meaning, consequently 'neg' is also added to their POS-tags (i.e. *min* | v.cop.neg and *med* | v.neg see below).

## Verbs

In principle ten verb tags are recognized, each of which has a verbal noun equivalent (i.e. suffixed with *-pa* or *-ba*). The four possible stems of Tibetan verbs are distinguished: present (v.pres), past (v.past), future (v.fut), imperative (v.imp). This distinction is made when it is morphologically marked (e.g. *gsod*, *bsad*, *gsad*, *sod* 'kill') and when it is contextually recoverable on the basis of sandhi phenomena (e.g. *gsol lo* = present, *gsol to* = past, *gsol cig* = imperative). Whenever it is not possible to distinguish stems on these two grounds the distinction is not made and the tag is simply 'v'. A separate tag (v.cop) is used for copulas such as *yin*, *lags*, and *mchis*.

An addition tag (v.aux) marks auxiliary verbs, i.e. verbs that appear after a finite verb, e.g. *rgyal-po chen-po khyod kyi lus la mar-me ston btsugs te mchod-pa byed nus na chos bstan-par byaḥo* 'O great king, if thou art able to make an offering, erecting a thousand butter lamps on thine body, then I shall teach the dharma' (vol. 74, page 131a). Elements falling into this class should not have stem inflection and are distinct from converbs. A verb from among this class is recognized as an auxiliary verb even if the verb which it governs is omitted, e.g. *rgyal-po ñid bzeñs te / srañ gi nañ du ḥgro-ba r gzas-pa las ñams kyis ma nus te* 'The king tried to raise himself and was about to go inside of the scale, but because of weakness was not able' (vol. 74, page 137a). In this sentence the last clause could be expanded to be *ḥgro ma nus te* 'he was not able to go'.

In order to represent reduplicated verbs, we introduce a special tag for the second element (v.redup). In our corpus so far, this element is always a verbal noun. Thus, in the phrase *ñin-thog skyel skyel-ba las* 'while he was gathering fruit', the first *skyel* is tagged as a present finite verb and the *skyel-ba* is tagged as a reduplicated verbal noun. There is no need to distinguish the stem in the reduplicated syllable because it is always an exact copy of the preceding syllable.

Finally, because negation is inherent to the meaning of the two verbs *min* and *med* negation 'neg' is also added to their POS-tags (i.e. *min* | v.cop.neg and *med* | v.neg).

## Affixes

As mentioned, we treat grammatical affixes as separate words; case markers are distinguished from converbs. This distinction may in fact be unnecessary and therefore unwise in those cases where cases and converbs are homophonous, but is nonetheless a prudent course of action.

Each case marker and converbial marker is distinguished with a separate part-of-speech tag. Naturally, phonologically predictable allomorphs (e.g. *-gi*, *-gyi*, *-kyi*, *-hi*, *-yi*) is brought under the same part-of-speech tag. The absolutive case, since it is zero-marked, will not be tagged or in any other way marked.

#### Cases

Table 4 presents the Tibetan case markers. The absolutive case is unmarked and is consequently untagged. This list of cases is based on morphosyntax and varies in several ways from the traditional reckoning of Tibetan case (cf. Hill 2012).

#### Converbs

Table 5 presents the Tibetan converbs. There is a great deal of overlap with the case markers. When an ambiguous morpheme, e.g. *-tu* is suffixed to a noun then it is tagged as a case marker and when it is suffixed to a verb it is tagged as a converb. Whenever a case marker is homophonous with a converb, we maintain the name of the case marker also for the converb; this practice implies nothing about how these converbs are used.<sup>11</sup>

Case name	Form	Abbreviation
Ablative	<i>-las</i>	case.abl
Agentive	<i>-gis</i> , <i>-gyis</i> , <i>-kyis</i> , <i>-s</i>	case.agn
Allative	<i>-la</i>	case.all
Associative	<i>-da</i> ᳚	case.ass
Comparative	<i>-bas</i> , <i>-pas</i>	case.comp
Elicative	<i>-nas</i>	case.ela
Genitive	<i>-gi</i> , <i>-gyi</i> , <i>-kyi</i> , <i>-᳚i</i>	case.gen
Locative	<i>-na</i>	case.loc
Terminative	<i>-tu</i> , <i>-du</i> , <i>-ru</i> , <i>-su</i> , <i>-r</i>	case.term

Table 4: Tibetan case markers

<sup>11</sup> What we call the 'final' converb, is a marker of finiteness, and thus is in no way what would normally be called a 'converb' in conventional linguistics. Nonetheless, since this item is a post verbal affix comparable to the others, it is convenient to label it analogously.

Case name	Form	Abbreviation
Agentive	<i>-gis, -gyis, -kyis, -s</i>	cv.agn
Allative	<i>-la</i>	cv.all
Elicative	<i>-nas</i>	cv.ela
Final	<i>-ho, -to, etc.</i>	cv.fin
Genitive	<i>-gi, -gyi, -kyi, -hi</i>	cv.gen
Imperative	<i>-ciñ, -ziñ, -síñ</i>	cv.imp
Imperfective	<i>-cig, -zig, -síg</i>	cv.impf
Locative	<i>-na</i>	cv.loc
Question	<i>-ham, -tam, etc.</i>	cv.ques
Semi-final	<i>-ste, -de, -te</i>	cv.sem
Terminative	<i>-tu, -du, -ru, -su, -r</i>	cv.term

Table 5 : Tibetan converbs

The general policy of allowing only case markers after nouns and only converbs after verbs is violated in two cases. We rarely analyse the genitive case marker as appearing directly appended to a verb stem. For example, *hgyur gyi mi* in the sentence *bdag-cag gi pha-ma dan hdra-ba khyod me-don du mtshon na / hbañs hdi dag thams-cad la mgon-skyabs dan / gnas med-par hgyur gyi mi hgah tsam gyi phyir / hbañs thams-cad kyi dpuñ-gñen med-par ma mdzad* 'If you, who are like our parents, jump into the fire pit, all these subjects will be some mere men without a protector or place, because of that do not extinguish the refuge of all subjects!' (vol. 74, page 134b), and *son gi phyir* in the sentence *rgyal-pos de skad ces thos nas rab tu khros te / mdañ gzu blañs nas rgyal-po ñid lag dar te khyeñu la hphañs nas mdañ hphangs pa khyeñu lam son gi phyir yañ rgyal-poñi druñ du lhuñ ño* 'When the king heard that, he became very angry. Taking up a bow and arrow, the king himself drew back his hand and shot at the person. The arrow that he show after the path the person had taken landed in front of the king' (vol. 74, page 146b). In the second exception, when *-na* means 'when' after a verbal noun we tag it as a converb rather than a case marker. For example, *skyes-pa na* in the sentence *miñi nañ du skyes-pa na / hdod-pas chog mi ses-pas gcig la gcig htshé zññ gnod-par gyur to //* 'When born among men, because of desire and discontent they hurt and harm one another' (vol. 74, page 134b).

### Clitics

The remaining POS-tags we use for a variety of heterogeneous word classes all of which appear to be clitics. Consequently, we begin these POS-tags with the letter 'cl'.

#### Focus clitics (cl.focus)

This tag is used for *ni* and *kyañ*.

#### Quotative clitic (cl.quot)

This tag is used for the quotative clitic with the allomorphs *ces*, *žes*, and *šes*. This clitic sometimes appears with the nominative suffix *-pa*; we see no need to distinguish these forms in our tag set.

#### The clitic *tsam* (cl.tsam)

This tag is used for *tsam*, which to us appears to have sui generis syntactic behaviour.

#### Punctuation (punc)

While a *tsheg* is considered part of the preceding syllable, all other punctuation marks are tagged as punctuation. So far we have encountered |, ĩ, ||, ༄།, and ||||.

### Software Infrastructure

A preliminary web site has been created for the early stages of the project<sup>12</sup>, and a first batch of materials has been uploaded, including the Tibetan and Himalayan Library's digital version of the Derge Kangyur, a set of texts kindly provided to us by the Tibetan Buddhist Resource Centre, and the Otani Tibetan E-Texts.<sup>13</sup> Together, these texts constitute well over 20 million syllables of Classical Tibetan, ample fodder for our initial experiments. Our corpus system is being built in Drupal, a PHP-based open-source content management system widely adopted within academic and archival communities.

### Pages

Tibetan texts are batch imported into our system from XML files that we receive from our providers, with each page of text corresponding to a single Drupal node. To the basic page and metadata fields, we add new fields holding the results of our hand

12 The URL, subject to change: <http://larkpie.net/tibetancorpus>

13 <http://web.otani.ac.jp/cr/twrp/project/otet/index.html>

annotations: for example, page text segmented into words, and page text tagged by part-of-speech. Pages with part-of-speech tags can be edited online or exported for offline editing.

Our use of pages as a basic organizing principle for Classical Tibetan texts enables us to more easily integrate our work with the systems of our data providers. For example, the THL presents the Derge Kangyur in a paged interface, with overlays of scanned pecha manuscript pages where available. We could ignore page breaks during tokenization and part-of-speech tagging, and then re-align the tagged texts to our providers' source files later, but by remaining faithful from the start to the organizing principles established by prior projects, we guard against the possibility that alignment, if delayed, would be deferred and ultimately left undone, thereby lessening the impact of our final products.

The main challenge presented by a page-driven structure is the scribal practice that pages are free to break after any syllable, whether it ends a word or not. Therefore words frequently do span across pages. This challenge has already been addressed: in addition to storing the original page text, the THL also stores a "page transition" including the "sentence" which spans the previous page and the start of the page in question, provided that page begins in the middle of a sentence. Adopting this approach prevents the appearance of orphaned syllables in our tagged texts, and also enables phrase queries to be executed across pages.

#### Datatags

To facilitate analysis and evaluation, we organize data on the website into "datatags". A *datatag* is a set of data which shares a common chain of analysis. Each datatag may reference a different word-segmenter or a different part-of-speech tagger, and in this way we can directly compare the performance of multiple models applied to the same data. For example, at the moment we are comparing the POS-tagging performance of Taku Kudo's CRF++ tagger<sup>14</sup> against that of Helmut Schmid's TreeTagger.<sup>15</sup> We are also examining the impact on POS-tagging performance of adding a simple rule-based tagger based on regular expressions search and replace operations over neighboring words and tags.

Our project is developing Java-based NLP tools for Tibetan, which are updated regularly on GitHub.<sup>16</sup> Drupal datatags divide the corpus into sub-parts that are passed through customizable natural

---

14 <http://crfpp.googlecode.com/svn/trunk/doc/index.html>

15 <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

16 <https://github.com/tibetan-nlp>

language processing chains. The NLP components are plugged into SOLR, the open-source Lucene-based search platform that plays nicely with Drupal.<sup>17</sup> Specifically, they are implemented as “update processor factories” that add, remove, or modify fields before sending a document (page) to the SOLR search index. These update processor factories are specified in SOLR's configuration file (solrconfig.xml), but a factory's settings depends on parameters passed with the document by the datatag.

Some of the NLP components are for general purpose use, not necessarily tied to Tibetan language. For example, the following SOLR configuration fragment instructs the CRF++ tagger to POS-tag the “twxs\_pos\_guessed” field using the supplied language model.

```
<processor
class="org.soas.solr.update.processor.CrfppTaggerUPF">
  <str name="datatags">im_field_datatags</str>
  <str name="path">ss_pos_path</str>
  <str name="model">pos</str>
  <str name="guess">twxs_pos_guess</str>
  <str name="uuidField">uuid</str>
  <str name="folds">01,23,45,67,89,ab,cd,ef</str>
  <str name="delimitOutput"> </str>
  <str name="tagDelimiter">|</str>
</processor>
```

The following fragment compares the guessed tags to those produced by hand-annotation, scoring accuracy as well as generated a field with the errors highlighted:

```
<processor
class="org.soas.solr.update.processor.TagScoreUPF">
  <str name="datatags">im_field_datatags</str>
  <str name="words">ts_field_pos_bo</str>
  <str name="guess">twxs_pos_guess</str>
  <str name="tokenCount">is_pos</str>
  <str name="guessCount">is_pos_guess</str>
  <str name="correctCount">is_pos_correct</str>
  <str name="errorField">twxs_pos_error</str>
  <str name="errorE">twxhtm_pos_error</str>
  <str name="errorTag">strong</str>
</processor>
```

---

17 For SOLR, see <http://lucene.apache.org/solr/>. For Drupal and Solr integration, see <http://drupal.org/project/apachesolr>.

Another factory tags a text using a lexicon instead of a statistical model. So, given a lexicon file containing a list of words with their possible tags (where dashes mark that we are ignoring lemmatization):

```
སླེབ་          v.fut -      v.past -
སླེབ་པ་       n.v.fut -    n.v.past -
```

Then we can use the fragment below to assign a list of possible tags to each word of a text:

```
<processor
class="org.soas.solr.update.processor.LexiconTaggerUPF">
  <str name="fieldRegex">twxs_possible_tags.*</str>
  <str name="lexicon">lexica/2013-02-
26_lexicon.txt</str>
  <str name="tagDelimiter">|</str>
  <str name="delimitOutput"></str>
</processor>
```

In addition to developing general purpose NLP components, we are also exploiting and extending existing Java libraries for Tibetan language processing to create a range of update processor factories for manipulating Tibetan syllables, words, and part-of-speech tags.

#### Rules

One of these Tibetan-specific components is a rule-based tagger that takes the output of the lexicon tagger and then eliminates impossible or highly unlikely tag sequences.

```
<processor
class="org.thdl.tib.solr.SimpleRuleTaggerUPF">
  <str name="fieldRegex">twxs_possible_pos.*</str>
  <str name="rules">lexica/2013-03-05_rules.txt</str>
</processor>
```

Here, the referenced document contains a list of rules formulated as regular expressions over neighboring words and tags.

We include rules to handle the homonymy of indefinite determiners and imperative converbs:

```
# ཅིག་()/ཞིག་()/ཤིག་()
```



# If a word 'w' tagged with a hypothesized POS-tag [v.pres] is followed by ཅིག(), ཞིག(), or འིག()

# and is preceded by མ' then delete all other hypothesized POS-tags.

```
((?:^|\s)མ'\|\S+)\s+(\S+)\|\|S*\[v\.pres\]\S*\s+((?:ཅིག|ཞིག|འིག)?)\| >
$1 $2|[v.pres] $3|
```

# If a word 'w' tagged with a hypothesized POS-tag [v.imp] is followed by ཅིག(), ཞིག(), or འིག()

# (and is not preceded by མ') then delete all other hypothesized POS-tags.

```
((?:^|\s)(?!མ)\S+)\s+(\S+)\|\|S*\[v\.imp\]\S*\s+((?:ཅིག|ཞིག|འིག)?)\| >
$1 $2|[v.imp] $3|
```

Additional rules, applied in the order below, help to distinguish between the previously mentioned inconveniently homonymous words *mi* 'person' and *mi* 'not'.

# མི'

# If མི' is followed by a word with the hypothesized tags [v.pres], [v.fut],

# [n.v.pres] or [n.v.fut], then assign tag [neg] to the word མི'.

```
((?:^|\s)མི'\|\S+)\s+(\S+\|((?:v\.pres|v\.fut|n\.v\.pres|n\.v\.fut)\|\|S*) > $1|[neg] $2
```

# If མི' follows a genitive, then assign tag [n] to it.

```
((?:^|\s)((?:འི|ཀྱི|གྱི)\|\S+)\s+(མི?)\|\S+ > $1 $2|[n]
```

# If མི' precedes མེདམ' then assign the tag [n] to it.

```
((?:^|\s)མི'\|\S+)\s+(མེདམ?)\| > $1|[n] $2|
```

# If མི'/མ' is followed by any one of the words རེ, འདི, འདི, རྣམས, དག, ཐམས་ཅད,

# ཀུན, ཞིག, འགའ, and གཞན, then delete the [neg] tag.

```
((?:^|\s)((?:མི|མ')\|\|S*)\|[neg\]\|S*\s+(((?:རེ|འདི|འདི|རྣམས|དག|ཐམས་ཅད|ཀུན|ཞིག|འག|གཞན)?)\| > $1|$2$3 $4|
```

We use the output of the rule-based tagger in various ways. First, it informs and improves occasionally inconsistent hand-tagging. Second, it assists in the rapid tagging of new material. And third, it

narrows the possibilities that need to be considered by our statistical models, thereby improving automated tagging performance.

#### Evaluation

Datatags are divided into folds. Upon ingestion into the system, each page of data is assigned a random (but unique) 32 digit hexadecimal code called its Universal Unique Identifier.<sup>18</sup> We divide the data into 8 folds based on the UUID prefix: '01' refers to data points whose UUID begins with '0' or '1', and so on up to 'ef'. Each fold is evaluated against models trained on the entire dataset, minus the fold. From the datatags page, you can download the '01' training data, which includes the entire datatag except those data points that begin with '0' or '1'. Models are built from this training data and then tested against the '01' fold. The process is repeated for the other folds. This is known as (8-fold) cross validation. Our interface therefore does not distinguish between 'training' and 'test' data; for every page, each datatag's performance can be checked against the human-annotated baseline.

As mentioned above, we are currently comparing and optimizing the performance of both the CRF++ tagger and the TreeTagger. On our (admittedly homogenous) *Mdzans blun* pilot data, the TreeTagger is currently leading the race with 98.1% part-of-speech tagging accuracy.

#### Future developments

The tagging score above assumes that our Tibetan text has been accurately broken into words. However, since Tibetan script does not mark word boundaries, a Tibetan part-of-speech tagger can be no better than the tokenizer that precedes it. This serves to highlight the critical importance of word segmentation to Tibetan NLP. Following Huidan et al (2011), we are experimenting with re-casting Tibetan word segmentation as a syllable tagging problem, with each syllable in search of an appropriate word-internal position label. For example, the only syllable of a monosyllabic word is tagged with 'S' for "single syllable", and the first, middle, and end syllables of multisyllabic words are tagged with "B", "M", and "E", respectively. This work is still in its early stages.

We will also soon be presented with further challenges when we unleash our taggers on other texts, as well as distinct historical periods and genres within our overall corpus.

---

18 See [http://en.wikipedia.org/wiki/Universally\\_unique\\_identifier](http://en.wikipedia.org/wiki/Universally_unique_identifier) for general discussion and <http://drupal.org/project/uuid> for the Drupal implementation that we are using.

## Conclusion

No Tibetan dictionary has been compiled which benefits from the advances in corpus linguistics which have revolutionized the lexicography of better studied languages. The compilation of a dictionary is always a major undertaking; a part-of-speech tagged corpus reduces the work and improves the outcome. Consequently, we call upon those who are considering starting a new Tibetan dictionary to invest their energy in the creation of the digital resources that will ensure a higher quality of dictionaries than would otherwise be possible. Although it is currently still in its beginning stages, 'Tibetan in Digital Communication' at SOAS promises to make a contribution in this direction.

## Appendix 1. Alphabetical list of POS tags

- Adj: ཚེན་པོ་, མང་པོ་, དམ་པ་, ཚེན་པོ་, རིང་པོ་, རྒྱང་ཏུ་, ལྷ་མ་, ལྷུང་, ལྷ་བུ་, གང་, དབྱུང་པོ་, མང་པོ་, མཚོན་པོ་, གཅིག་པ་, གཅིག་པ་, ཉམ་ཐག་པ་, ཐ་རྒྱུད་, བཟང་པོ་, འཕགས་པ་, རིང་, གཞི་པོ་, གཡམ་པ་, དན་པ་, རོམ་ཚར་, དུ་མ་, བཟང་པོ་, མཚོག་, མདོངས་པ་, རིང་པོ་, རྒྱན་མོ་, གཅིག་པུ་, གཏི་ལྷག་ཅན་, གཏི་ལྷག་ཅན་, གཅང་མ་, གཞི་པོ་, གཞོན་ནུ་, བྲངས་མེད་, རྒྱུ་ཟད་, ཚེབ་ཅན་, ཉག་མ་, ཉམ་རྒྱུད་, ཐ་མལ་པ་, ཐ་མལ་པ་, དཀར་མུ་, དཀར་མུ་, དབྱུང་པོ་, དབྱུང་པོ་, རོན་མེད་, རོན་མོ་, ལྷི་མ་, ཞར་མ་, ཡ་མཚན་, རོ་བ་, རོ་བ་, རིང་པོ་, རིན་པོ་ཆེ་, རོན་པོ་, རོན་པོ་, ལྷམ་མེར་, ལྷ་མ
- adv.dir: ཡན་རྒྱུན་, ལྷིན་ཅད་, ལྷོན་ཅད་, མན་ཅད་, ཡན་ཅད་, ལྷན་ཅད་
- adv.intense: རབ་(ཏུ་), ལྷིན་(ཏུ་)
- adv.proclausal: དེ་(ནས་), དེ་(ལྷོ་), གཡམ་(ཏི་), འོ་(ན་), ཡང་(ན་), འོན་(ཏི་)
- adv.temp: ལྷོན་, ད་, དེང་, མདང་, གཤོད་, དུང་, ལྷི་ལྷིན་, ལྷི་ལྷོ་, མང་
- case.abl: ལས་, ལས
- case.agn: ས་, ལྷིས་, ལྷིས་, ལྷིས་, ས་, ལྷིས་, ལྷིས་
- case.all: ལ་, ལ
- case.ass: དང་, དང་<sup>19</sup>
- case.comp: བས་, བས་
- case.ela: ལས་, ལས
- case.gen: འི་, ལྷི་, ལྷི་, ལྷི་, ལྷིས་<sup>20</sup>

19 Unicode distinguishes two types of *tsheg*, which is the reason for this word to occur twice.

20 In the text itself -gis is a misprint for -gi.

- case.loc: ན་, ན
- case.term: འི་, འི་, འི་, འི་, འི་, འི་, འི་,
- cl.emph: ཉིད་, ལོ་ན་, རེ་རེ་, ལོ་
- cl.focus: ཡང་, ནི་, ཀྱང་, ནི་, ཅང་, མུར་ཡང་, འང་
- cl.indef: ཞིག་, ཅིག་, ཤིག་, ཞིག་, ཞིག་, ཤིག་
- cl.quot: ཞེས་, ཅེས་, ཞེ་, ཅེས་པ་, ཅེས་བ་, ཞེས་པ་
- cl.tsam: ཅོམ་, ལྟོད་, ཅོམ
- cv.agn: ཀྱིས་, ལྱིས་, ཀྱིས་, ཤིས་, ལྱིས་
- cv.all: ལ་, ལ་
- cv.ela: ནས་, ནས
- cv.fin: མོ་, ཏོ་, འོ་, ཏོ་, རོ་, འོ་, ནོ་, ཏོ་, ལོ་, ཏོ་, ཤོ་, ཤོ་, ལོ་, རོ་, ཏོ་, རོ་, ཏོ་, རོ་, ཏོ་
- cv.gen: ཀྱི་, ལྱི་, ཀྱི་, ཤི་, ཤི་, ལྱི་
- cv.imp: ཤིག་, ཅིག་, ཅིག་, ཤིག་
- cv.impf: ཞེང་, ཅེང་, ཤིང་, ཞེང་
- cv.loc: ན་, ན
- cv.ques: དམ་, འམ་, ནམ་, དམ་, འམ་, མམ་, མམ་, མམ་, མམ་, མམ་, མམ་, མམ་, མམ་, མམ་, མམ་
- cv.sem: ཉི་, ཉི་, ཉི་, ཉི་, ཉི་
- cv.term: འི་, འི་, འི་, འི་, འི་, འི་
- d.dem: དེ་, འདི་, དེ་, འདི་
- d.det: གཞན་, འབའ་, ཡ་རེ་
- d.emph: ཉིད་, ལོ་ན་, རེ་རེ་, ལོ་
- d.indef: ཞིག་, ཅིག་, ཤིག་, ཞིག་, ཞིག་, ཤིག་
- d.plural: དག་, ཐམས་ཅད་, ཀྱན་, མཐའ་, ལྷོ་ལྗོངས་, ཡོངས་, འུང་བུད་, འགའ་, ཐམས་ཅད་, དུམ་, མོ་ཙོག་
- dunno: ཙོག་ཙོག་, གཏན་, གམལ་ཀ་, ལྱིན་, བྱངས་, མཁྱེན་, ལྟ་
- n: ལུས་, དུས་, བཅོམ་ལྷན་འདས་, ཆོས་, ལྱལ་པོ་, ལྱལ་པོ་, སངས་རྒྱས་, མི་, སེམས་, བཀའ་, ལྱལ་བུ་, ལྟ་, ལུ་, འཛིག་ཉེན་, འཁོར་, སེམས་ཅན་, མོ་བུད་, མོ་ན་པོ་, ལྟ་, ལྱལ་, བཙུན་མོ་, ལྱལ་པོ་, རྒྱུ་, སྒྲིམ་, འབྲས་བུ་, ལུས་, འོངས་ལྟོད་, ལྟོད་, གང་ཉེན་, མ་མ་, བསོད་ནམས་, རིགས་, ལྷག་པམུལ་, སྟོག་, ལྲིམ་, འབྲས་བུ་, གནམ་, མྱིང་, ལུང་མ་, ཐམ་མོ་, དཤེ་འདུན་, དཤེ་ལྟོད་, དོན་, ཚུང་, འཛམ་བུ་, ཡི་, རིན་པོ་ཆེ་, ནམ་པ་, ལས་, ལེ་བུ་, སྐྱ་, ལྲིས་བདག་, ལུགས་བཅེས་, བསྐྱེད་པ་, ཀྱན་དགའ་རབ་, གཏམ་, ཆར་, མེག་མ་, དགའ་

བཅའ་མཁའ་, ཚེག་, འཕྲོ་བ་, རྣམ་པ་, སྤྲུལ་, གཏོན་, བུངས་, བོད་ཆེར་, དབྱ་བཅོམ་པ་, བ་མ་, ལྷོ་ལོ་, བཅོན་འགྲུལ་, འུ་མོ་, བམ་ཟེ་, མེ་ཏོག་, ཚེ་, ཞབས་, ཡུན་, ལོ་, ཤ་, རྒྱ་, རྩོན་པ་, ཁྲིམས་, འུ་, བུང་རྒྱལ་, ཟླ་, མི་, མིང་, ཚེགས་, ཞག་, འཕྲོད་པ་, ཡོན་ཏན་, རྒྱ་མཚོ་, རྒྱལ་མཚན་, རྒྱལ་ཚབ་, རྩེ་, རྒྱབས་, རྩོན་ལམ་, རྩུ་, ལ་, ལག་, ཚེས་མོས་, ཉིན་, དཔེ་, དབང་པོ་, བཟོ་ལ་, རྩོགས་, མགོ་པོ་, མཐ་, མཚན་, ཞལ་ཟམ་, རུམ་པ་, རྒྱ་མཚོ་, རྩོ་, རྩུན་, རྩུང་མ་, དེོས་པོ་, དབང་, བཀའ་འདྲིན་, བམོད་རྩོམས་, བུང་རྒྱལ་མེམས་དཔལ་, བུང་གཞུགས་, བམ་ཟེ་, རྩོན་པོ་, མཚིམ་, མེ་, ཚོང་པ་, འཁོར་བ་, འོད་, ཡལ་, ཡོ་བུད་, རྒྱལ་པུ་, ལན་, ལྷ་རྩེ་, རྩོ་, ལྷག་མོ་, ལྷན་, ལམས་, གཏོ་སྤྲུག་, གདན་, གཏོ་སྤྲུག་, རྩུང་པོ་ཚེ་, ཚོ་འཇུག་, བམ་, བག་, ལྷགས་, དལྱ་, དགེ་བཅོམ་, དལྱགས་, དལྱལ་བ་, དེ་བཞིན་གཤེགས་པ་, དོང་ཚེ་, དོར་, བཀའ་པ་, རུམ་མོ་, ལྷག་པོན་, པོ་ཉེ་, བཀ་, བལེགས་, བཅོན་མོ་, བུང་རྒྱལ་མེམས་དཔལ་འཇུག་, མ་, མ་སྤྲང་, མང་འབྲུད་, མའོན་སྤུལ་, མཐུ་, མཐོ་མེས་, མའུན་, མའོག་, མིག་, མེ་, མེ་དོར་, རྒྱལ་ཁྲིམས་, ཞེ་སྤང་, ཟམ་, ཟླ་, འདོད་རྒྱལས་, འབངས་, འཕྲོང་པ་, ཡི་དགས་, ཡུམ་, རང་མངས་རྒྱལ་, རིན་པོ་ཚེ་, རིམ་, རྒྱུན་, རྒྱང་གཤགས་, རྒྱ་, རྒྱན་, རི་, འག་, འག་པ་, འམ་, རྩུག་མ་, འིད་, རྒྱད་, རྒྱེ་པོ་, རྩི་ལྷོ་, རྩ་, རྩུན་སྤྲུ་, རྩི་པོ་, རྩང་, རྩམ་, རྩམ་མོ་, རྩུ་, ལ་, ལ་དོག་, ལ་སྤྲུ་, ལམ་, རྩུལ་མ་, འིད་, རྒྱད་, རྩི་པོ་, རྩི་ལྷོ་, རྩ་, རྩུན་སྤྲུ་, རྩི་པོ་, རྩང་, རྩམ་, རྩམ་མོ་, རྩུ་, རྩུང་, རྩུམིག་, ཚོངས་, ཉམ་, ཉིམ་, ཉིན་ཞག་, ཉེག་ནས་, ཉེས་པ་, བམས་, བེ་ཚོམ་, དཀོན་མཚོག་, དགེ་པ་, དབྱ་བཅོམ་པ་, དཔག་, དཔུང་གཉེན་, དཔེ་, དབང་པོ་, དལྱ་, དལྱལ་ཕོངས་, དལྱིབས་, དམ་, དམ་བཅའ་པ་, དམ་བཅའ་པ་, དམ་གཤམ་, དུམ་ལུ་, དོར་, བན་ཏན་, བམ་, བ་, བན་, ལུ་མོ་, ལྷག་དར་, སྤམ་, བདུད་, བམ་པོ་, བཅོན་ཏུག་, བཅོན་པ་, བཟའ་བ་, བར་ཚན་, བར་སྤང་, བཅོད་, བཅུན་, བསྐལ་བ་, འུ་མོ་, འུ་མོ་, འུམ་མོ་, བུད་, ཟློ་, མགོ་, མཚེམ་, མཐའ་, མང་པ་, མདོ་, མཚན་མཁའན་, མའ་མེ་, ཞལ་, འཁོར་པོ་, འགམ་, འགམ་ལུ་, འཕྲིང་པོ་, ཡིག་, ཡི་དམ་, ཡི་ཤེས་, རབ་བསྐྱེད་པ་, རབ་, རམ་ལུ་, རལ་གྱི་, རལ་གྱི་, རི་, རུམ་ལུ་, རེ་, ཀང་པ་, རྒྱང་མ་, རྒྱན་, རྒྱལ་མཚན་, རྩོད་, རྩོང་ལུ་, འག་མཐེལ་, རྩོངས་, རྩུས་, རྩེ་མིག་, རྩུང་, རྩུན་, རྩུང་བཟེད་, འིད་ཐོག་, འིད་ན་, རྩུན་པ་, རྩིང་, རྩིང་ལྷོ་, རྩིང་ལྷོ་, རྩིས་པ་, རྩུག་པོ་, རྩིག་པ་, རྩུ་, ཀ་པ་, ཀུའུ་, ཀུན་ཏུ་རྒྱུ་, ཀུན་ཏུ་རྒྱུ་བ་, ཀུན་ཏུ་རྒྱུས་, ལ་ཟམ་, ལང་པ་, ལང་པ་, ལང་བཟངས་, ལལ་, ལས་མ་, ལུ་རྩུང་, ལུ་པེ་, ལོ་མོ་, ལོམ་ཐབ་, ལོ་ཤུག་, ལོ་སྤྲུན་, ལུ་, ལུས་, གཅན་གཟན་, གཉེན་ཟླ་, གཏུན་, གདན་པ་, གདན་པ་, གཏུག་པ་, གཏུགས་, གན་, གནས་, གནས་ཁང་, གནས་པ་, གཞུ་, གཞོངས་, གཟི་བཞིན་, གཞུགས་བུད་, གཡམ་, གཡམ་, གཡོག་འཁོར་, གཡོན་, གམོད་ནམས་, གྲི་, གྲོགས་, གྲོགས་པོ་, རྩོ་པ་, དན་མེས་, དན་མོང་, དན་རྒྱུགས་, དལ་, དུད་འཕྲོ་, དེས་, དོ་, དོ་འཇུམ་, ཚག་ཚག་, ཚང་, ཚབ་, ཚབ་མིག་, ཚབ་པོམ་, ཚབ་སློབ་, ཚབ་སློབ་, རྩུ་པོ་, རྩུའན་, ཚོག་, ཚོད་, ཚོ་སྤྲུགས་, ཚོས་, ཉམས་, ཉུག་ཚམ་, བ་རྒྱང་, བད་, བལ་མོ་, རྩུགས་ཐེ་ཚེན་པོ་, བེ་ཚོམ་, བོ་, དལྱིལ་འཁོར་, དལྱེ་རྩུང་, དེོས་པོ་, དེོས་སྤྲིག་, དཔེ་བུད་, དབང་ཐང་, དབང་པོ་, དལྱལ་ཕོངས་པ་, དཔེན་པ་, དཔག་པ་, དཔེགས་ལུ་, དུག་, དུད་འཕྲོ་, དུམ་ལུ་, དེ་བཞིན་གཤེགས་པ་, དེ་བཞིན་གཤེགས་པ་, དེད་དཔོན་, དང་ཕོང་, དི་བསྐྱེད་, དི་མ་, དི་མ་, དིན་, བན་, ལུ་པོ་, ལུ་པོ་, ལུ་བཅོམས་, ལེུག་མིང་, ལོར་དལྱིག་, ལོར་བདག་, བགས་, བདམ་, བས་མོ་, རུན་སྤུམ་ཚོགས་པ་, ལྷོ་, ལྷོ་བཞིན་, ལྷོ་མི་, ལྷོན་ལོ་ལོ་, ལྷག་པོན་དལྱིག་, སྤྲུག་པོན་, སྤྲུག་པོན་, སྤྲུ་, སྤྲུ་, སྤྲུ་, སྤྲུ་, སྤྲུ་མི་ཤེས་, སྤྲུ་མི་ཤེས་, སྤྲུ་,

བག་ཆགས་, བག་མ, བག་མེད་, བགོ་བ་, བང་མཚོན་, བཏུང་བ་, བན་སྐྱང་, བརྗོན་, བཟང་མོ་, བཟང་ཤིང་,  
 བར་, བརྟན་, བརྟུན་པ་, བསྐྱ, བཤད་གཏེ་, བསྟེན་བཀའ་, བུ་ཚུང་, བུ་ཅེ་, བུ་ཚོ་, བུ་, བྱང་རྒྱལ་སོམ་ས་,  
 བྱང་ལྗོགས་, བྱས་མ་པ་, བྱིང་ལིང་གང་ལི་, བྱིང་ལིང་གང་ལི་, བྱུང་པ་, བློ་ཅན་, མཉམ་ས་པ་, མཉོན་, མཉོན་སྐྱུབ་ས་,  
 མགྲོན་པ་, མཛེན་, མཚན་ལུང་, མཚེས་བུང་, མཚེད་, མཚོན་རྟེན་, མདངས་, མདོ་, མཚོན་, མཚོན་ཆ་,  
 མཛེད་ས་སྐྱུན་, མཛེས་མ, མངས་ཤེ་, མི་ལྟོགས་, མི་ཤ, མུན་ཁང་, མེ་སྐྱུང་, མོ་ཁབ་, མོད་, ཚང་, ཚངས་རིས་,  
 ཚལ་པ་, ཚིག་རྒྱུབ་, ཚུལ་, ཚོར་བ་ས་, ཚོགས་, ཚལ་ཆེབ་, ཚིང་, ཚེག་ཅོད་པ་, ཚེ་སྐྱང་ཚོབ་, ཟན་, ཟབས་,  
 ཟལ་རྒྱལ་, ཟལ་, འམ་འོལ་, འཁོར་བ་, འཁོར་ལོ་, འདུ་ཤེས་, འཕྱིན་, འཕྱོང་བ་, འདྲ་ཟེང་, འདྲ་སྐྱུང་,  
 ཡིད་ཚོམ་པ་, ཡོན་བདག་, རབ་, རབ་, རིད་གས་, རིའོ་, རིར་བ་ལྷན་པོ་, རིགས་, རིན་, རིས་གྲོ་, རིས་པ་,  
 རིལ་ཤིང་, རིས་, རུམ་, རོ་, རོ་ལ་, རྟ་, རྟུང་, རྟུན་ཏུ་ལྷགས་པ་, ར་, རེའུ་, རོདོག་, རྟགས་,  
 རྟན་ཅིང་འབྲེལ་པར་འབྱུང་བ་, རུལ་, རེའུ་, རོ་རྗེ་, རོ་རྗེ་, མི་ལམ་, ཅ, ཅ་བ་, ཅིལ་ལོགས་, ཅ་, ཡུང་,  
 ཡུས་གཞུགས་, ཡེ་ལོ་, ཡེའུ་, ཡོང་བ་, ཡོངས་, ཡུགས་གཟེང་, ཡྱིང་ཀ, ཡུ་མ་ཡིན་, ཡུ་མ་རོམ་, ཡུ་མོ་, ཡུ་མོ་,  
 ཡུ་རྟེན་, ཡུན་, ཡུ་ལྗོགས་, ཤས་ཐབས་, ཤར་ཚོགས་, ཤིང་རོག་, ཤིང་རྟ, ཤུལ་, ཤེས་རབ་, ཤོག་ཤོག་,  
 ཤེར་སྐྱོ་, ཤོ་, ཤོ་ཆེས་, ཤྲུལ་ལྱུང་, ཤྲུལ་སྟོན་, ཤྲུ་ཚོ་, ཤྲུ་གནས་, ཤྲེས་ཤུ་, ཤྲོན་, ཤྲོབས་པ་, ཤྲོབ་, ཤྲོབ་,  
 ཤྲོ་སྐྱུང་, ཤྲུ་, ཤྲུ་ཅལ་, ཤྲུ་, ཤྲོན་མ་, ཤྲོས་, ཤྲོས་ཤུ་, ཤྲོ་ལོ་, ཤྲིང་པོ་, ཤྲིང་པཅེ་, ཤྲུའུང་, ཤྲལ་, ཤྲལ་སྐྱུག་,  
 ཤྲལ་སྐྱུགས་, ཤྲོན་པ་, ཤྲོན་མོ་, ཤྲེ་, ཤྲེ་, ཤྲལ་ཅེ་, ཤྲལ་ཚོ་, ཤྲལ་, ཤྲང་, ཤྲོས་, ཤྲུའུ་, ཤྲང་, ཤྲུའོ་,  
 ཤྲིན་བདག་, ཤྲུང་རིགས་, ཤྲུལ་, ཤས་པོ་, ཤས་པོ་, ཤིན་པོ་, ཤྲང་མ་, ཤོག་གཏོད་, ཤྲུང་བཞིན་པ་, ཤྲོབ་དོན་,  
 རང་བཞིན་

n.mass: གལེར་, ལྷ་, རོར་ལུ་, ཅུར་སྟེན་, ཟངས་, དུལ་, མཐོན་མཐིང་

n.prop: ཀུན་དགའ་པོ་, རྒྱལ་བྱེད་, བརྒྱ་བྱིན་, མགོན་མེད་ཟས་སྦྱིན་, ཀུན་དགའ་པོ་, མཉན་ཡོད་, གལེར་དཀྱིག་,  
 བྲ་རྟ་སྟེ་, གང་བླ་དར་, གསལ་རྒྱལ་, གལེར་སྟབ་, ཚངས་པ་, ལྷའི་མེ་རྟོག་, ལྷའི་རིན་ཆེན་, སེལི་,  
 སེམས་ཅན་ཆེན་པོ་, གང་བླ་ བྲ་རྟ་སྟེ་ སི་ལཱ་ཀཱ་, ལྷས་པ་, མཉན་ཏུ་ཡོད་པ་, ལྷ་ཆེན་པོ་, ཤིང་རྟ་ཆེན་པོ་,  
 རྒྱུད་མོས་, རྒྱ་ཆེན་པོ་, ཀུན་ཤིན་པ་ལི་, ལ་དོག་དམ་, ལ་དོག་དམ་པ་, གལེར་དཀྱིག་, དལ་བསོལ་པོ་, དལ་ས་,  
 དམ་གས་, དགའ་ལྷན་, དབང་སྐྱུག་ཆེན་པོ་, དམ་གས་, དེ་བ་བརྟ་, བསྐྱ་མི་བྲ་, བཀའ་ཤིས་མ་, སི་ལཱ་ཀཱ་,  
 ལཱ་ཀ་ལན་དགའ་, མ་ག་རྒྱ་, མ་ཤི་བྲུང་, མ་སྟེམ་དགའ་, ཚངས་པ་, ཚངས་པ་ལྷ་, ཚངས་པ་ས་བྱིན་, ཟས་གཏང་མ་,  
 ཟས་གཏང་མ་, འབར་ལི་, འདྲ་མ་, ལྷའི་རོ་རྗེ་, རྟལ་ཐོས་, རྟལ་པར་རྒྱལ་པ་, ཡེའུ་དུ་ཅི་, ཡེའུ་དུ་ཆུ་, ལྷ་རིལི་བྲ་,  
 ལྷ་མ་ལྷ་མཛོམ་, ལྷ་དཔལ་པ་, མ་སྟེམ་དགའ་

n.rel: སྐད་, ཚེ་, ལྷུ་, བཞིན་, ལྷ་, ཅང་, ལྷལ་, བར་, ལྷུ་, ལྷ་བྲུ་, ལྷེང་, ལྷེ་, ལྷུང་, འོག་, ཚེ་, འདྲུ་, དབུས་, ལྷུ་,  
 ལྷུ་བཞིན་, འབྲལ་

n.v: གསོལ་པ་, ཐོས་པ་, མཛོན་པ་, ཐོབ་པ་, མཚོན་པ་, གསུངས་པ་, དགའ་པ་, འདོད་པ་, ལོགས་པ་, ལྷག་པ་, དགའ་པ་,  
 ལྷན་པ་, ལྷུང་པ་, བདེ་པ་, བསྐོབ་, ཟབ་, རྒྱགས་པ་, ལོན་པ་, ལྷགས་པ་, སེས་པ་, གསོལ་པ་, བཅད་པ་, བཅས་པ་,  
 མཐོབ་, མཐོང་པ་, འདོད་པ་, ཡོད་པ་, ལྷགས་པ་, གསོལ་པ་, ཐར་པ་, དགོབ་, བདེ་པ་, བཟང་པ་, མཐོབ་, འདྲུ་པ་,  
 རིགས་པ་, ལྷན་པ་, གསོགས་པ་, གསོལ་བ་, ཚོགས་པ་, ལྷགས་པ་, དད་པ་, དད་པ་, བཅས་པ་, བཛེན་པ་,









### Bibliography

- Beckwith, Christopher I. (2001). Review of Goldstein 2001. *Anthropological Linguistics* 43.3: 395-398.
- Bray, John (2008). "Missionaries, officials and the making of the 1826 Dictionary of the Bhotanta, or Boutan Language." *Zentralasiatische Studien* 37: 33-75.
- Chandra, Lokesh (1958-1961). *Tibetan-Sanskrit dictionary, based on a closed comparative study of Sanskrit originals and Tibetan translations of several texts*. New Delhi: International Academy of Indian Culture.
- Chandra, Lokesh (1992-1994). *Tibetan-Sanskrit dictionary. Supplementary volumes*. New Delhi: International Academy of Indian Culture and Aditya Prakashan.
- Chandra, Lokesh (2007). *Sanskrit-Tibetan dictionary: being the reverse of the 19 volumes of the Tibetan-Sanskrit dictionary*. New Delhi: International Academy of Indian Culture and Aditya Prakashan.
- Chimpa, Lama, Bimalendra Kumar, and Jampa Samten, eds. (2011). *Meghadūta: critical edition with Sanskrit and Tibetan index*. New Delhi: Aditya Prakashan.
- Ejima, Yasunori, et al. (1985-1993). *Index to the Saddharmapuṇḍarīkasūtra: Sanskrit, Tibetan, Chinese*. Tokyo: Hotoke no Sekaisha,
- Francke, Herbert et al. (2005-). *Wörterbuch der tibetischen Schriftsprache*. Munich: Verlag der Bayerischen Akademie der Wissenschaften.
- Goldstein, Melvyn C., ed. (2001). *A New Tibetan English Dictionary of Modern Tibetan*. University of California Press.
- Hill, Nathan W. (2007). "Personalpronomina in der Lebensbeschreibung des Mi la ras pa, Kapitel III." *Zentralasiatische Studien* 36: 277-287.
- Hill, Nathan W. (2010). "Personal Pronouns in Old Tibetan." *Journal Asiatique* 298. 2: 549-571.
- Hill, Nathan W. (2012). "Tibetan -las, -nas, and -bas." *Cahiers de Linguistique—Asie Orientale* 41.1: 3-38.
- Hirakawa, Akira (1973-1978). *Index to the Abhidharmakośabhāṣya*. Tokyo: Daizō Shuppan.
- Huidan Liu, Minghua Nuo, Longlong Ma, Jian Wu, and Yeping He (2011). "Tibetan Word Segmentation as Syllable Tagging Using Conditional Random Field." 25th Pacific Asia Conference on Language, Information and Computation: 168–177.

- Inagaki, Hisao (1984). *A tri-lingual glossary of the Sukhāvaiṣyūha sūtras: indexes to the Larger and Smaller Sukhāvaiṣyūha sūtras*. Kyoto: Nagata Bunshodo.
- Jäschke, Heinrich August (1881). *Tibetan English Dictionary*. London: Unger Brothers.
- Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell (2004). "The Sketch Engine." Geoffrey Williams and Sandra Vessier, eds. *Proceedings of the eleventh Euralex International Congress*. Lorient: Université de Bretagne-Sud. 105-116.
- Kilgarriff, Adam and Iztok Kosem. "Corpus tools for lexicographers." *Electronic Lexicography*. Sylviane Granger and Magali Paquot, eds. Oxford: Oxford University Press. 31-55.
- Maurer, Petra and Johannes Schneider (2007). "Neues Datenbanksystem für das Wörterbuch der tibetischen Schriftsprache." *Akademie Aktuell* 22.3: 23.
- McGrath, Bill (2008). "Tibetan Dictionaries." <http://www.thlib.org/reference/dictionaries/tibetan-dictionary/dictionary-biblio.php> (accessed, 5 March 2013)
- Ñag dbañ tshul khrims (1997). *Brda dkrol gser gyi me long*. Beijing: Mi rigs dpe skrun khang.
- Nagao, Gajin (1958-1961). *Index to the Mahāyāna-sūtralaṅkāra*. Tokyo: Nihon Gakujutsu Shinkōkai.
- Nagao, Gajin (1994). *An index to Asaṅga's Mahāyānaśāstra*. Tokyo: The International Institute for Buddhist Studies.
- Negi, J. S. (1993-2004). *Tibetan-Sanskrit dictionary*. Sarnath: Dictionary Unit, Central Institute of Higher Tibetan Studies.
- Obermiller, Eugène (1970). *Indices verborum Sanskrit-Tibetan and Tibetan-Sanskrit to the Nyāyabindu of Dharmakīrti and the Nyāyabinduṭīka of Dharmottara*. Osnabrück: Biblio-Verlag.
- Schmid, Helmut (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees." *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.
- Schneider, Johannes and Petra Maurer (2012). "Ein Wörterbuch des Tibetischen." *Akademie Aktuell* 40.1: 50-51.
- Simon, Walter (1964). "Tibetan Lexicography and Etymological Research." *Transactions of the Philological Society* 63.1: 85-107.
- Suzuki, Daisetz Teitaro (2000). *An index to the Lankavatara sutra (Nanjio edition): Sanskrit-Chinese-Tibetan, Chinese-Sanskrit, and Tibetan-Sanskrit*. New Delhi: Munshiram Manoharlal Publishers.
- Uebach, Helga and Jampa L. Panglung (1998). "The Project "Dictionary of Written Tibetan" : An Introduction." *Lexicography in the Indian and Buddhist cultural field*. Boris L. Oguibénine, ed. Munich: Kommission für Zentralasiatische

- Studien, Bayerische Akademie der Wissenschaften. 149-163.
- Walter, Michael (2006). "A bibliography of Tibetan dictionaries." *Bibliographies of Mongolian, Manchu-Tungus, and Tibetan dictionaries*. Hartmut Walravens, ed. Wiesbaden: Harrassowitz. 174-235.
- Weller, Friedrich (1933). *Index to the Tibetan translation of the Kāçyapaparivarta*. Cambridge: Harvard-Yenching Institute.
- Yamaguchi, Susumu (1974). *Index to the Prasannapadā Madhyamaka-vṛtti*. Kyoto: Heirakuji-Shoten.
- Yokoyama Kōitsu (1996). *Index to the Yogācārabhūmi, Chinese-Sanskrit-Tibetan*. Tokyo: Sankibō Busshorin.
- Zeisler, Bettina (2002). "The development of temporal coding in Tibetan: some suggestions for a functional internal reconstruction. (1): Unexpected use of the 'imperative' stem in Old Tibetan and Themchen (Amdo Tibetan)." *Tibet, Past and Present*. Henk Blezer, ed. Leiden: Brill. 441-453.
- Zhang Yisun (1985). *Bod rgya tshig mdzod chen mo*. Beijing: Mi rigs dpe skrun kañ.

