himalayan linguistics

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

Himalayan Linguistics

Algorithmic description of the decomposition and checking of a Classical Tibetan written syllable

Elie Roux

Buddhist Digital Resource Center

Hélios Hildt

Buddhist Digital Resource Center

ABSTRACT

This document presents our research on the the correct formation of a Classical Tibetan syllable. It was triggered by attempts at defining the boundaries of well-formed syllables in Classical Tibetan for spell checking purposes.

Formalizing the formation of the syllable led us to inspect the small differences among grammar books, both in Western and Tibetan language. We then checked these differences against the Tibetan dictionaries we consider reliable, and also against the Kangyur.

Our inquiry finally led us to study the way to decompose a syllable, discussing the ambiguous cases, as well as the formation of the Dzongkha syllable.

KEYWORDS

Tibetan, spell checking, NLP

This is a contribution from *Himalayan Linguistics, Vol. 17(1): 50-66.*
ISSN 1544-7502

Himalayan Linguistics, Vol. 17(1). © Himalayan Linguistics 2018
ISSN 1544-7502

Algorithmic description of the decomposition and checking of a Classical Tibetan written syllable

Elie Roux
Buddhist Digital Resource Center

Hélios Hildt
Buddhist Digital Resource Center

1 Introduction

This document presents our research on the orthography of a Classical Tibetan syllable. The study of this question was triggered by attempts at defining the boundaries of well-formed written syllables in Classical Tibetan for NLP purposes. What we are trying to assess is the extend of the possibilities of which a modern Tibetan person would think as valid if they wanted to invent a new Classical Tibetan word. This is a necessary first step to be able to discriminate valid Classical Tibetan from other forms such as transliterated Sanskrit, or just mistakes. Of course, the syllable level is not enough for a full feature spell checker; but we hope our work will provide a basis for future work in this area, as well as a complete description of the various corner cases of syllable construction. This research has already been useful, allowing us to create

- a basic spell checker for Classical Tibetan using the *hunspell* library
- rules for collation of Classical Tibetan

Formalizing the formation of the syllable led us to inspect the small differences between grammar books, both in Western and Tibetan language[1]. We then checked these differences against some chosen Tibetan dictionaries[2] and the Derge Kangyur[3].

Our inquiry finally led us to study the way to decompose a syllable, discussing the ambiguous cases, as well as the formation of the Dzongkha syllable.

[1] Our main readings in Western languages were (Tournadre and Dorje, 2005) and (Beyer, 1992), but we also consulted (Bacot, 1946) and (Das, 1915); the Tibetan sources were (སི་ཏུ་པཎ་ཆེན་ཆོས་ཀྱི་འབྱུང་གནས།, XVIIIth c.) and (ཚེ་ཏན་ཞབས་དྲུང་།, 2003). The bibliography used in this document can be downloaded on https://github.com/eroux/tibliography/ .
[2] The dictionaries we consulted are (Dorje et al., 2003), (སྨོན་ལམ།, 2016), (དུང་དཀར་བློ་བཟང་འཕྲིན་ལས།, 2002), and in a lesser extent (Yisun, 1985), (ཆོས་སྐད་རྫོང་ཁ་ཚིག་མཛོད་ཆེན་མོ།, 2010) and (Negi, 1993). We also consulted (Hill, 2010).
[3] (ཆོས་ཀྱི་འབྱུང་གནས།, 1721) , input by Esukhia, https://github.com/Esukhia/derge-kangyur

2 Elements of a Tibetan syllable

2.1 Description and vocabulary

2.1.1 Basic description

For the description of the different elements of a Tibetan syllable, we will take the vocabulary used in (Tournadre and Dorje, 2005). As an example, decomposing the syllable བརྒྱངས་ results in radical letter ག, prefix ར, superscribed ར, subscribed ཡ, vowel accent ◌ྱ, first suffix ང and second suffix ས. Among these six categories, only the radical letter is mandatory, all the others are optional[4]. This constitutes a good framework for the description of the elements of a syllable but lacks a few elements.

2.1.2 Missing elements

First, some syllables (such as དབགས་) contain a wasur (◌ྭ). These are traditionally considered as subscribed letters, with the ability to combine with other subscribed (as in ཐྭ). This last case breaks our initial description because it requires two optional subscribed letters instead of one, and the second can only be a wasur. Our proposal is to treat the wasur separately from the subscribed letters in order to keep things simple. So དབགས་ would be decomposed as radical letter ད, wasur, first suffix ག, second suffix ས; and ཐྭ as radical letterག, subscribed ར and wasur.

A second problem is the ཨ "suffix" (as in ལེའུ་, "chapter", /leu/, 2 syllables): it not considered as a suffix in traditional grammars (that are usually completely silent about it), and does not have normal suffix properties. We will call it a *special suffix*. At most one can appear in a syllable.

The third missing feature is affixed particles. These are appended to syllables with no suffixes or replace suffix ཨ. For instance མའི་ is decomposed as radical letter མ and affixed particle ཡི. Two affixed particles can even be combined, as in མའིའོ་.

In order to describe the different elements, we will call *final part* what is written on the right side of the radical letter; *main stack* the radical and everything written above and below except vowel accents (so subscribed, superscribed letters and wasur).

We also would like to introduce the notion of *root*, which is constituted of everything preceding the vowel accent or final part. This notion will come handy in some parts of this document, and is crucial for collation purposes.

2.1.3 Formalization

Now we can propose a formalization of the elements of the syllable with some symbols coming from the regular expressions: "?" means an element that can either be omitted or appear once, and "|" simply means an exclusive "or":

syllable = [root][vowel]?[final part]?
root = [prefix]?[main-stack]
main-stack = [superscribed]?[radical][subscribed]?[wasur]?
final-part = ([special-suffix][affixed-particles]?) | [suffixes] | [affixed-particles]
affixed-particles = [affixed-particle][affixed-particle]?

[4] For the sake of simplicity, we will not consider that the "a" vowel accent or the first suffix ཨ are implied when no vowel accent or first suffix is present.

suffixes = [first-suffix][second-suffix]?

Note that although this is the way a syllable is built, some syllables can have mutiple decomposition possibilities according to this scheme. For instance བར་ can be decomposed as either:

- radical བ + suffix ར (/bar/, "between")
- radical བ + affixed particle ར (either /war/ nominalizer+oblique mark, or /bar/ "cow"+oblique mark)

This schema also doesn't account for implied suffix ད: for instance if དཀའི་ did not have its affixed particle, it would be དཀའད་, and for some purposes it may also be useful to retain this information, but we chose the most simple schema for readability purposes.

2.2 Constraints on the construction of a syllable

In this part we will study the different possibilities of construction of these elements, first for each element taken independently, then the constraints of their relations.

2.2.1 Simple constraints

All consulted sources agree on all the constraints on the different elements except the special suffix. These constraints are:

- prefix can be གྲྀདབམ or ཨ
- first suffix can be གངདྷབམཨརལ or ས
- second suffix can be ས[5]
- superscribed can be ར ལ or ས
- subscribed can be ྭ ྱ or ྵ
- radical letter can be ཀཁགངཅཆཇཉཏཐདནཔཕབམཙཚཛཝཞཟའཡརལཤཥསཧ or ཨ
- vowel accent can be ི ུ ེ or ོ
- wasur is ྲྀ
- affixed particle can be འི འོ འང འམ ར or ས

When inspecting the different constraints on the relation between elements, we can see some consistent lists among all sources:

- ཀཁགངཇཉཏདནབམཙཚཛཞཟའཡཤཧཨ for superscribed ར + radical letter
- ཀཁགངཅཆཇཉཏདནཔཕབམཙཚཛཝཞཟའཡཤཧཨ for superscribed ལ + radical letter

[5] Old Tibetan also has ད.

- སྐ སྒ སྔ སྙ སྟ སྡ སྣ སྤ སྦ སྨ སྩ for superscribed ས + radical letter
- ཀྱ ཁྱ གྱ པྱ ཕྱ བྱ མྱ for radical letter + subscribed ཡ
- ཀླ གླ བླ ཟླ རླ སླ for radical letter + subscribed ལ
- རྐྱ རྒྱ རྨྱ for superscribed ར + radical letter + subscribed ཡ
- སྐྱ སྒྱ སྤྱ སྦྱ སྨྱ for superscribed ས + radical letter + subscribed ཡ

2.2.2 Methodological choices

But other constraints are inconsistent among sources. Before we expose the constraints we see as the most relevant, we must expose our methodological choices.

First, the language we deal with is what is usually called Classical Tibetan; and we take it in a modern perspective: XXIth century Classical Tibetan. This leads us to consider that some old texts contain errors even if they follow what the rules of Classical Tibetan were at their period.

The syllables we are inspecting are what is sometimes referred to as ཁྲིམས་མཐུན་གྱི་ཚེག་བར།, literally "legal syllables". While there is no clear cut definition of this concept, we exclude syllables appearing solely in 1. transliterated foreign languages, 2. proper nouns[6], 3. misspellings, 4. words in other languages such as Zhang zhung language (ཞང་ཞུང་སྐད་) or Dzongkha; but we chose to include 1. onomatopoeia (སྒྲ་སྐད་), 2. regional orthographies (ཡུལ་སྐད་), 3. words coming from Old Tibetan (བརྡ་རྙིང་), 4. words that are alternate spellings of other words. When a syllable comes from such a word, it will be suffixed by (OT) for Old Tibetan, (O) for onomatopoeia, (R) for regional words or (AS) for alternate spelling.

There is a whole spectrum of choices that can be taken for this kind of research, on one side of the spectrum we could have taken all the syllables in a purely Tibetan dictionary such as (Dorje et al., 2003) and consider that only these are legal syllables. On the other side we can take the most inclusive constraints on the formation of a legal syllable and state that all the combinations respecting these constraints are legal. We chose an in-between solution: quite strict constraints and listing exceptions. This has the inconvenience to imply some subjectivity in the words we include or exclude based on the supposed relevance of the sources, but we hope that the article will go in enough details to allow anyone to adapt this strategy to their own needs.

2.2.3 Constraints on the subscribed ར

The subscribed ར has the following possible combinations with radical letters in all the grammars we consulted: ཀྲ ཁྲ གྲ དྲ པྲ ཕྲ བྲ and སྲ. But some sources list additional possibilities: ཏྲ ཐྲ ནྲ ཧྲ and མྲ. Among these, མྲ and ཧྲ are extremely rare and appear only in the following words: ཨ་མྲ་ ("mango") and ཧྲིག་ཧྲིག་(O), so we consider the syllables མྲ་ and ཧྲིག་ as exceptions[7]. ནྲ is present in many words so we can count it as regular. ཏྲ is present only in ཏྲ་ ("ape"), ཏྲེས་སམ་ ("powder") and པ་ཏྲ་ (name of a particular

[6] Even the very common person names ཀརྨ་ and པདྨ་ are not considered "legal syllables", so we consider other proper nouns shouldn't. We do not consider the names of plants or lunar mansions as proper nouns when they do not come directly from a foreign language.

[7] An "exception" here means that no other syllable can be built with these stacks, for instance མྲན་ would not be considered legal.

design), so we consider the syllables ཀུ་ and ཀེས་ as exceptions. We have not been able to find any word with the main stack ཀྲ except proper nouns, so we don't consider it as a possibility at all.

For combinations with radical letter and superscribed ར, all sources indicate རྐ རྒ རྦ and རྟ. Some indicate རྩ and རྨ. རྨ is present only in rare words such as རྨོན་, རྨེབས་ (two lunar mansions), and རྨེ་ལྷ་ཞིག (OT, "disordered"); རྩ is only found in རྩ་ ("speak"), རྩོད་(OT, "speak") and derivatives: རྩོལ་, རྩལ་, རྩེལ་ རྩེད་ རྩངས་,[8] so we consider only these exceptions as valid for རྩ and རྨ.

One of the sources not citing རྩ nor རྨ as likely to be used in new words is (Beyer, 1992), invoking phonological reasons.

2.2.4 Constraints on the wasur

Wasurs explanation is even more different among sources. All the source for main stacks with wasurs give at least ཀ ཁ ག ཅ ད ཧ ཨ ཤ ཏ ཧྲ and སྟ. It is also possible to find ཏྲ, ཥ and ས in some others. Our research showed that the word ད་ (contraction of དེ་བ་) can also be found in dictionaries.

The main problem with these descriptions is that they are too loose: among the many possibilities of syllables containing a wasur, so few exist that it seems easier to see them as exceptions. Also, the wasur does not seem to be used in the creation of new words[9], so we chose to treat the syllable with wasurs as exceptions and list them. We have been able to list the following 25 syllables: [10]

ཀ་(O) ཀའི་(O) ཁ་ ག་ ཐ་ ཅ་ དོ་(OT) དངས་ དགས་ ད་ ཐུ་ ཧྲ་ ཚ་ ཚོས་ ཨ་ ཤ་ ཉ་ སོ་ བས་ བསོ་(last three: O, AS for བས་) ཏ་ ཏཱ་(in ཐེ་ཏཱ་, OT for ཐྲུང་རྒྱ་, "honey") ཏཱས་.

2.2.5 Constraints on prefix+radical

The following rules on prefixes are found in all grammars:

- **ག** can only be prefix of ཅ ཇ ཉ ད ན ཙ ཨ ཤ ཧ or ས
- **ད** of ཀ ག ང ཟ ཏ or ཧ
- **བ** of ཀ ག ཅ ཉ ད ཙ ཨ ཤ or ས
- **མ** of ཁ ག ང ཆ ཇ ཉ ཐ ད ན ཚ or ཥ
- **འ** of ཁ ག ཆ ཇ ཐ ད ཟ ཏ ཚ or ཥ

But (Tournadre and Dorje, 2005) gives other possibilities for prefix ཏ combined with ང ཇ ཉ ན and ཥ. This difference can easily be explained: in these cases, ཏ cannot be prefix of the radical letter when there is no superscribed (ex: ཏངའ་ is impossible), it only can when a superscribed is present (ex: ཏརྟ་).

[8] These possibilities are found in the various sources of (Hill, 2010).

[9] (Tsering, 2006) does not list any word with wasur.

[10] We did not include ཀྱ་, བས་ (in proper nouns), རང་, རེ་, དུང་ (Zhang zhung language), ད་, ཀོཏ་, ཀཱན་, ཀྱང་, ཤྲང་ (Chinese), དོ་གས་ (given only in (མཁས་སྒོང་རྩོ་རྡེ་དབང་ཕྱུག, 1979)), དུ, རྒྱ nor སྐ (Dzongkha).

2.2.6 Constraints on root

The main additional constraint found in traditional grammar is that prefixes ག ད ཨ and འ cannot appear when a superscribed letter is present.

But (ཞི་ཀྲ་པཎ་ཆེན་ཆོས་ཀྱི་འཇྱུང་གནས།, XVIIIth c.) goes in more details and gives a more specific list of possible combinations of roots11. Leaving aside the wasurs and the naked radical letters, the given possibilities are the following:

ཀྱ ཀྲ ཁྲ དཀྱ དཀྲ དཏྲ དགྱ དགྲ དབྱ དབྲ ཏྲ ཐྲ ཇྱ ཇྲ དཀྱ དཀྲ དབྱ དབྲ དཔྱ དཔྲ དཧྲ དགྱ དགྲ དབྱ དབྲ དཔྱ དཔྲ ཧྲ ཀྱ ཁྲ དཀྱ དཀྲ དབྱ དབྲ དཔྱ དཔྲ ཧྲ ཨཀ ཨཁ ཨག ཨགྲ ཨང ཨཅ ཨཇ ཨཉ ཨཏ ཨད ཨདྲ ཨན ཨཔ ཨཕ ཨབ ཨབྲ ཨམ ཨཙ ཨཛ ཨཝ ཨཞ ཨཟ ཨའ ཨཡ ཨར ཨལ ཨཤ ཨས ཨཧ ཨཨ ཀཀ ཀཁ ཀག ཀགྲ ཀང ཀཅ ཀཇ ཀཉ ཀཏ ཀད ཀདྲ ཀན ཀཔ ཀཕ ཀབ ཀབྲ ཀམ ཀཙ ཀཛ ཀཝ ཀཞ ཀཟ ཀའ ཀཡ ཀར ཀལ ཀཤ ཀས ཀཧ ཀཨ ཁཀ ཁཁ ཁག ཁགྲ ཁང ཁཅ ཁཇ ཁཉ ཁཏ ཁད ཁདྲ ཁན ཁཔ ཁཕ ཁབ ཁབྲ ཁམ ཁཙ ཁཛ ཁཝ ཁཞ ཁཟ ཁའ ཁཡ ཁར ཁལ ཁཤ ཁས ཁཧ ཁཨ གཀ གཁ གག གགྲ གང གཅ གཇ གཉ གཏ གད གདྲ གན གཔ གཕ གབ གབྲ གམ གཙ གཛ གཝ གཞ གཟ གའ གཡ གར གལ གཤ གས གཧ གཨ ངཀ ངཁ ངག ངགྲ ངང ངཅ ངཇ ངཉ ངཏ ངད ངདྲ ངན ངཔ ངཕ ངབ ངབྲ ངམ ངཙ ངཛ ངཝ ངཞ ངཟ ངའ ངཡ ངར ངལ ངཤ ངས ངཧ ངཨ ཅཀ ཅཁ ཅག ཅགྲ ཅང ཅཅ ཅཇ ཅཉ ཅཏ ཅད ཅདྲ ཅན ཅཔ ཅཕ ཅབ ཅབྲ ཅམ ཅཙ ཅཛ ཅཝ ཅཞ ཅཟ ཅའ ཅཡ ཅར ཅལ ཅཤ ཅས ཅཧ ཅཨ ཇཀ ཇཁ ཇག ཇགྲ ཇང ཇཅ ཇཇ ཇཉ ཇཏ ཇད ཇདྲ ཇན ཇཔ ཇཕ ཇབ ཇབྲ ཇམ ཇཙ ཇཛ ཇཝ ཇཞ ཇཟ ཇའ ཇཡ ཇར ཇལ ཇཤ ཇས ཇཧ ཇཨ ཉཀ ཉཁ ཉག ཉགྲ ཉང ཉཅ ཉཇ ཉཉ ཉཏ ཉད ཉདྲ ཉན ཉཔ ཉཕ ཉབ ཉབྲ ཉམ ཉཙ ཉཛ ཉཝ ཉཞ ཉཟ ཉའ ཉཡ ཉར ཉལ ཉཤ ཉས ཉཧ ཉཨ ཏཀ ཏཁ ཏག ཏགྲ ཏང ཏཅ ཏཇ ཏཉ ཏཏ ཏད ཏདྲ ཏན ཏཔ ཏཕ ཏབ ཏབྲ ཏམ ཏཙ ཏཛ ཏཝ ཏཞ ཏཟ ཏའ ཏཡ ཏར ཏལ ཏཤ ཏས ཏཧ ཏཨ ཐཀ ཐཁ ཐག ཐགྲ ཐང ཐཅ ཐཇ ཐཉ ཐཏ ཐད ཐདྲ ཐན ཐཔ ཐཕ ཐབ ཐབྲ ཐམ ཐཙ ཐཛ ཐཝ ཐཞ ཐཟ ཐའ ཐཡ ཐར ཐལ ཐཤ ཐས ཐཧ ཐཨ ཇཀ ཇཁ ཇག ཇགྲ ཇང ཇཅ ཇཇ ཇཉ ཇཏ ཇད ཇདྲ ཇན ཇཔ ཇཕ ཇབ ཇབྲ ཇམ ཇཙ ཇཛ ཇཝ ཇཞ ཇཟ ཇའ ཇཡ ཇར ཇལ ཇཤ ཇས ཇཧ ཇཨ ཉཀ ཉཁ ཉག ཉགྲ ཉང ཉཅ ཉཇ ཉཉ ཉཏ ཉད ཉདྲ ཉན ཉཔ ཉཕ ཉབ ཉབྲ ཉམ ཉཙ ཉཛ ཉཝ ཉཞ ཉཟ ཉའ ཉཡ ཉར ཉལ ཉཤ ཉས ཉཧ ཉཨ ཏཀ ཏཁ ཏག ཏགྲ ཏང ཏཅ ཏཇ ཏཉ ཏཏ ཏད ཏདྲ ཏན ཏཔ ཏཕ ཏབ ཏབྲ ཏམ ཏཙ ཏཛ ཏཝ ཏཞ ཏཟ ཏའ ཏཡ ཏར ཏལ ཏཤ ཏས ཏཧ ཏཨ ཐཀ ཐཁ ཐག ཐགྲ ཐང ཐཅ ཐཇ ཐཉ ཐཏ ཐད ཐདྲ ཐན ཐཔ ཐཕ ཐབ ཐབྲ ཐམ ཐཙ ཐཛ ཐཝ ཐཞ ཐཟ ཐའ ཐཡ ཐར ཐལ ཐཤ ཐས ཐཧ ཐཨ

(ཚེ་ཏན་ཞབས་དྲུང་།, 2003) also gives two exceptions of syllables that do not follow these constraints: བཟླ་ and མདོན་. We were able to find མདོན་ as an abbreviation of མདོ་དྲན་ ("to remember a sutra") in (བཙུ་ཡིག་གསེར་ཀྱི་ཨ་ལོང་། / A Handbook of Abbreviations, 2011), but did not find བཟླ་ in the sources we consulted12, so we do not consider them valid. We also found བཟློ་ cited in (Kirtivajra, 1982), but we are not sure of its relevance.

As noted earlier, our proposal is to also to remove ཀ, ཟ, སྣ, ཟ and སྣ from the list and only consider the exceptions listed in previous paragraphs.

It would be possible to also treat the following as exceptions instead of rules:

- བརྐྱ found only in བརྐྱངས་ and བརྐྱངས། (both: "to extend")
- བཟླ found only in བཟླག་, བཟླགས་ (both: "to read out loud"), བཟླྭད་ and བཟླྭངས་ (both: "to wear")
- བཙ found only in བཙད་ ("to smile") and བཙལ་ ("to get past")

But most of these pertain to conjugation and restricting these possibilities may not be wise, as new verbs using these may appear in the future.

2.2.7 Constraints on vowels and suffixes

The most documented rule for suffixes is that the second suffix ས can only be found after the first suffixes ག ང ད or ཨ.

We have not been able to find any constraint on vowels or first suffixes other than འ.

For first suffix འ, the rule given in (Beyer, 1992) can be easily verified; we choose to formalize it as follows: first suffix འ can only appear in syllables with one prefix, no superscribed, no subscribed and no vowel accent. We can thus easily list 48 possible syllables with suffix འ:

[11] Although the list is not given directly, it is quite straightforward to infer, and it is given more directly in (Das, 1915) and (ཚེ་ཏན་ཞབས་དྲུང་།, 2003).

[12] Although they have been found in (ས་ཆེན་ཀུན་དགའ་སྙིང་པོ, 1092), which we did not have time to fully explore for this article.

དཀའ་ བཀའ་ མགའ་ ལགའ་ དགའ་ བགའ་ མགའ་ འགའ་ དངའ་ མངའ་ གཆའ་ བཆའ་ མཇའ་ ཉའ་ གཉའ་ བཏའ་ མཆའ་ འཆའ་ མཐའ་ འཐའ་ གཏའ་ བདའ་ མནའ་ འནའ་ གནའ་ མཁའ་ དབའ་ འབའ་ དབའ་ འབའ་ དམའ་ གཙའ་ བཙའ་ མཚའ་ འཚའ་ མཛའ་ འཛའ་ གཟའ་ བཟའ་ གཡའ་ གཤའ་ བཤའ་ གསའ་ བསའ་

The only legal syllables with suffix འ not in this list are དཉེའ་(OT, AS for དཉེ་, “unclean”) and བརྟའ་(OT, AS for བརྟ་, “sign”)13. We can also find the onomatopoeia ཨེའ་(O), in which the suffix འ cannot be replaced by an affixed particle, which could maybe be pushed forward as in the (extremely unlikely) ཨེའའི་.

2.2.8 Constraints on affixed particles

Affixed particles can only appear when there is no suffix or first suffix ད, in which case the first suffix ད is replaced by the affixed particle. Affixed particles are mutually exclusive except འི་ that can be combined (ཡིའོ)[14]. This combination is only documented in (Beyer, 1992) but we were able to find this pattern around 200 times in (ཚོས་ཀྱི་འབྱུང་གནས།, 1721). No other combination seems possible.

The pronunciation of affixed particles and special suffxes depend on the dialects and the context, sometimes adding one or several phonetic syllables. This complex topic will not be discussed here as it has no impact on the matter at hand.

2.2.9 Special suffixes

Special suffixes are very rarely documented, we did not find any documentation in Tibetan language. The special suffix is similar to a suffix, but can be followed by an affixed particle (without being removed).

As this feature is not common, we choose to list the syllables containing a special suffix ཨ:

ཟླའོ་ སྐྱེའོ་ ཀཱའུ་ ཀིའུ་ ཀེའུ་ ཁེའུ་ ཁོའུ་ ཁྱིའུ་ ཁྲེའུ་ ཁྲོའུ་(OT) གཱའུ་ གིའུ་ གེའུ་ གོའུ་ གྱེའུ་ གྲེའུ་ གྲོའུ་ འགིའུ་ ཀེའུ་ ཁཱའུ་ ཁིའུ་ ཁྱིའུ་ ཁྱེའུ་ ཁྲེའུ་ ཅེའུ་ ཆེའུ་ ཇེའུ་ ཉེའུ་ ཏཱའུ་ ཏིའུ་ ཏེའུ་ ཐེའུ་ ཐོའུ་ ཐྱེའུ་ ཐྲེའུ་(OT) ཉེའུ་ ཉོའུ་ ཏཱའུ་ ཏིའུ་ ཏེའུ་ ཐེའུ་ དཱའུ་ དིའུ་ དེའུ་ དོའུ་ དྱེའུ་ དྲེའུ་ དྲོའུ་ ནཱའུ་ ནིའུ་ ནེའུ་ ནོའུ་ ནྱེའུ་ ནྲེའུ་ ནྲོའུ་ པཱའུ་ པིའུ་ པེའུ་ པོའུ་ པྱེའུ་ པྲེའུ་ པྲོའུ་ ཕཱའུ་ ཕིའུ་ ཕེའུ་ ཕོའུ་ ཕྱེའུ་ ཕྲེའུ་ ཕྲོའུ་ བཱའུ་ བིའུ་ བེའུ་ བོའུ་ བྱེའུ་ བྲེའུ་ བྲོའུ་ མཱའུ་ མིའུ་ མེའུ་ མོའུ་ མྱེའུ་ མྲེའུ་ མྲོའུ་ ཙཱའུ་ ཙིའུ་ ཙེའུ་ ཙོའུ་ ཙྱེའུ་ ཙྲེའུ་ ཙྲོའུ་ ཚཱའུ་ ཚིའུ་ ཚེའུ་ ཚོའུ་ ཚྱེའུ་ ཚྲེའུ་ ཚྲོའུ་ ཛཱའུ་ ཛིའུ་ ཛེའུ་ ཛོའུ་ ཛྱེའུ་ ཛྲེའུ་ ཛྲོའུ་ ཝཱའུ་ ཝིའུ་ ཝེའུ་ ཝོའུ་ ཞྱེའུ་ ཞྲེའུ་ ཞྲོའུ་ ཟཱའུ་ ཟིའུ་ ཟེའུ་ ཟོའུ་ ཟྱེའུ་ ཟྲེའུ་ ཟྲོའུ་ འཱའུ་ འིའུ་ འེའུ་ འོའུ་ འྱེའུ་ འྲེའུ་ འྲོའུ་ ཡཱའུ་ ཡིའུ་ ཡེའུ་ ཡོའུ་ ཡྱེའུ་ ཡྲེའུ་ ཡྲོའུ་ རཱའུ་ རིའུ་ རེའུ་ རོའུ་ རྱེའུ་ རྲེའུ་ རྲོའུ་ ལཱའུ་ ལིའུ་ ལེའུ་ ལོའུ་ ལྱེའུ་ ལྲེའུ་ ལྲོའུ་ ཤཱའུ་ ཤིའུ་ ཤེའུ་ ཤོའུ་ ཤྱེའུ་ ཤྲེའུ་ ཤྲོའུ་ སཱའུ་ སིའུ་ སེའུ་ སོའུ་ སྱེའུ་ སྲེའུ་ སྲོའུ་ ཧཱའུ་ ཧིའུ་ ཧེའུ་ ཧོའུ་ ཧྱེའུ་ ཧྲེའུ་ ཧྲོའུ་ ཨཱའུ་ ཨིའུ་ ཨེའུ་ ཨོའུ་ ཨྱེའུ་ ཨྲེའུ་ ཨྲོའུ་(R) ཤྲེའུ་(OT) ཤྱིའུ་ ཤྲེའུ་ ཤྲོའུ་ ཡཱའུ་ ཡེའུ་ འཡེའུ་(OT) ལེའུ་ ཐྲའུ་ ཐྲིའུ་ ཐྲེའུ་(AS for ཐྲིའུ་) ཐྲའུ་ ཐེའུ་ ཐོའུ་ འཐེའུ་(OT) སྲྭའུ་(OT) མིའུ་ སྲིའུ་(OT, AS for མིའུ་) རྨིའུ་ (OT, AS for སྨིའུ་) སྨིའུ་(OT) ཙོའུ་ གཙོའུ་ རྩིའུ་ རྩེའུ་ ཚོའུ་ ཚོའུ་ མཆོའུ་ མཆོའུ་ རྩིའུ་ རྩེའུ་ གཤྲའུ་ ཟེའུ་ ཡེའུ་(OT) གཡིའུ་ རེའུ་ ལའུ་(OT) ལིའུ་(AS for ལེའུ་ in· ལེའུ་ཐོ་) ལེའུ་ ཤའུ་ ཤེའུ་ སིའུ་ སེའུ་ སྲིའུ་ སྲེའུ་ བསེའུ་ ཨའུ་.[15]

Other special suffixes appear exceptionally, they are འ in ཟླའོ་ (R, AS for ཟླ་བ་, “buckwheat”), སྐྱེའོ་ (OT, “conversation”) and ཀྱཻའི་ (O).[16]

[13] We did not consider བརྟའ་ given in (མཁས་སྒོོ་ངོ་རྩེ་དབང་རྒྱུག, 1979) as this source is not considered reliable.

[14] Note that this allows to build written syllables with the 4 vowels such as ལེའུའིའོ་ (“the one of the chapter”), pronounced [lewy:jo:] (3 syllables) by one informant with a Standard Tibetan accent.

[15] It doesn’t include ཀྱའུ་ ཁྱའུ་ རྒྱའུ་ ཤྱའུ་ དུའུ་ ཝེའུ་ རྒྱའུ་ ཁྱའུ་ དུའུ་ ཉེའུ་ རྒྱའུ་ ཅོའུ་ (found in Chinese words), གོའུ་ འགོའུ་ (found in Sanskrit words), ནོའུ་ (Mongolian), ཀྱའུ་ ཅོའུ་ གཅོའུ་ nor དྲེའུ་ (found in proper names).

[16] (Tournadre and Dorje, 2005) also cites འ but doesn’t give any example and we were not able to find any.

3 Decomposing a syllable

3.1 Finding the main stack

What we propose here is to find the main stack in any valid syllable. Once the main stack is found, a set of rules described in a later section can be applied to find the different elements. Some rules found in (Tournadre and Dorje, 2005) describe a general way for doing that but are not precise enough to be implemented. For instance they would fail in cases such as བཞི་, ལེའུ་, བཅམ་, etc. Based on the information we gathered from our research, we propose to apply the following rules in this particular order[17]:

1. if the syllable contains a subscript, superscript or wasur then the main stack is what contains it (ex: དབྱངས་ → ཡ)
2. if a letter other than ཨ carries a vowel then it is the main stack (གཱི་ → ཨཱ)
3. if a vowel is carried by ཨ and ཨ is the first letter then ཨ is the main stack (ཨོད་ → ཨ)
4. if a vowel accent is carried by an ཨ which is not the first letter, the main stack is before the first ཨ with a vowel[18] (མའིའོ་ → མ)

The following rules deal with syllables with no vowel, superscribed, subscribed nor wasur, only composed of letters that could be radical letters:

5. if the syllable has three or four letters and ends with ཅང or ཅམ, then the main stack is right before ཅང or ཅམ[19] (བཅམ་ → བ)
6. if the syllable is composed of only one letter this letter is the main stack
7. if the syllable contains two letters, then the first is the main stack (བར་ → བ)
8. if the syllable contains four letters, the main stack is the second letter (བཟབས་ → ཟ)

The following rules deal with syllables composed of a sequence of 3 simple letters:

9. if the final letter is not ས, then the main stack is the second letter (བཟའ་ → ཟ)
10. if the first letter cannot be a prefix to the second letter when it has no superscribed nor subscribed (see above for conditions on prefixes), then the main stack is the first letter (ཇགས་ → ཇ)
11. if ས cannot be second suffix after the second letter if it was a first suffix (meaning the second letter is not ག, ང, ད nor མ), then the main stack is the second letter (གཅོས་ → ཅ)

3.2 Ambiguous syllables

If the syllable doesn't match any of the above rules, then it is ambiguous. These syllables are those with no explicit vowel, no superscribed, no subscribed, no wasur, three letters, a final ས, the first letter can be a prefix to the second letter with no superscribed nor subscribed, and the second letter

[17] This means that the first rule the syllable will match will be the one determining the main stack. This has the advantage to be transcribed easily into computer code.

[18] This rule also works for Old Tibetan using ཪས. The only exception to this rule is the very unlikely མའའི་(O), but the legality of this syllable is not clear.

[19] This rule could be extended for Old Tibetan with ཅར.

is ཀ, ང, བ or མ. It is easy to find the 9 corresponding cases: མངས་, མགས་, དབས་, དངས་, དགས་, དམས་, བགས་, ཨབས་ and ཨགས་.

For each of these cases we have three possible structures:

1. ས is a second suffix (e.g. མངས་ can be decomposed as radical མ, suffix ང, second suffix ས, we note this decomposition མ|ངས)
2. ས is a suffix (e.g. མངས་ is radical ང, prefix མ, suffix ས, noted མང|ས)
3. ས is an affixed particle (e.g. མངས་ radical ང, prefix མ, suffix ཨ replaced by affixed particle ས, noted མང|ཨ+ས)

Case 2 and 3 have equivalent pronunciations but are important to distinguish because they will imply a different analysis in terms of lemmatization and part of speech tagging.

In order to decide where the main stack is, the only way is to take the decomposition with the highest probability according to our knowledge of the Tibetan language.

Let's review the different possibilities for the 9 cases, prefixing by * a form unattested in the dictionaries we consulted:

- **མངས་:** མ|ངས, *མང|ས, མང|ཨ+ས → ambiguous, **མ** as main stack is more intuitive to our Tibetan informants
- **ཨབས་:** *ཨ|བས, *ཨབ|ས, ཨབ|ཨ+ས → **ས** is the main stack
- **མགས་:** *མ|གས, *མག|ས, *མག|ཨ+ས → ambiguous, **མ** as main stack is more intuitive to our Tibetan informants
- **བགས་:** བ|གས, *བག|ས, *བག|ཨ+ས → **ས**
- **ཨགས་:** *ཨ|གས, ཨག|ས, ཨག|ཨ+ས → **ག**
- **དབས་:** *ད|བས, དབ|ས (OT), དབ|ཨ+ས → **ས**
- **དགས་:** *ད|གས, དག|ས, དག|ཨ+ས → **ག**
- **དངས་:** ད|ངས (misspelling of དྲངས་[20]), *དང|ས, *དང|ཨ+ས → **ང**
- **དམས་:** *ད|མས, དམ|ས, དམ|ཨ+ས → **མ**

3.3 Decomposition of the syllable

Once the main stack is found, the prefix, vowel accent and wasur are immediate to find. If a superscript or subscript is present, they can be immediately found with the rules exposed in "Simple constraints".

Suffixes can be immediately classified between special suffix, first suffix, second suffix and affixed particle; the exceptions are ས and ར, plus the syllables ཀྱིའོ, སྒོའོ, སྣེའོ. The latter are very rare and the context should make it obvious, but they are not decidable at syllable level.

[20] Attested in (Negi, 1993) (vol. 6) and (Duff, 2000).

The disambiguation of ས and ར can sometimes be done at syllable level according to the existence or absence of one version of the syllable; ex: in ཟས་, ས can only be affixed particle, because radical ཟ + suffix ས is nowhere to be found in dictionaries.

But many common cases are not decidable without the context. The most infamous example would be the very common བར་, in which we cannot decide if ར is suffix or affixed particle (see *2.1.3*). This is a very significant difficulty for some areas of NLP because this ambiguity makes phonetic transcription, particle inflection, etc. dependent of the context and thus difficult to achieve systematically. Note that this does not impact collation as both forms are strictly equivalent from this point of view.

We implemented this decomposition in a very simple software[21] and tested it against all the data produced for (Hildt, 2016), based on (ཚེ་ཏན་ཞབས་དྲུང་།, 2003). The only two differences were two different choices for ambiguous syllables, so the main rules succeed on 100% of the ~18000 syllables.

3.4 The Dzongkha syllable

Dzongkha syllables have the same general structure as the Tibetan syllable, with some variations we will detail here. The rules we describe come mostly from (གཞི་རིམ་རྫོང་ཁའི་བརྡ་གཞུང་།, 2010) and private discussions with members of the Dzongkha Development Commission (hereafter referenced as DDC) about a Dzongkha syllable spell checker for the *hunspell* library[22]. During these discussions, we received an official list of 48,721 words from the DDC, which have been our main resource for research.

3.4.1 Additional wasurs and ཝ forms

The word list we received contains the following wasurs and ཝ forms that have not been encountered in Classical Tibetan sources:

ཅེཝ་ ཛེཝ་ ཏཝ་ དཝ་ ནེཝ་ ཕོཝ་ ཙཝ་ ཚོཝ་ ཞེཝ་ རྣེཝ་ ཐེཝ་ ལྟེཝ་ སྐེཝ་ ཤེཝ་ སྦེཝ་ ཏཝ་ ཏེཝ་ གྭཝ་ དྭ་ ཟྭོ་ སྭ་

3.4.2 Additional affixed particles

In addition to the Classical Tibetan affixed particles, it is also possible to find the following particles in affixed form: པ པོ བ བོ མ མོ. All these particles are written at the end of a syllable in their usual form, except བ which is noted ཝ (as in དཝ་, not དབ་, which would introduce some ambiguity with the བ suffix). We were able to infer the following rules:

- no other affixed particle can be affixed after these particles (ex: རྒྱལཔོའི། is not correct)
- unlike other affixed particles, they do not replace the འ suffix (ex: མདའཔོ། is correct)
- possible combinations of suffix + particles are given in (གཞི་རིམ་རྫོང་ཁའི་བརྡ་གཞུང་།, 2010), we reproduce the table here:

[21] Available on https://github.com/eroux/tibetan-spellchecker/tree/master/examples.

[22] Available on https://github.com/eroux/hunspell-dz.

	པ	པོ	བ༼ཞི༽	བོ	མ/ མོ
ག	x	x			x
ང	x	x			x
ད	x	x	x		x
ན	x	x			x
བ	x				x
མ	x	x			
འ			x	x	x
ར	x	x	x		x
ལ	x	x	x		x
ས	x	x	x		x
-	x	x	x	x	x

At the time of the redaction of this article, some questions are still pending, like the possibility to affix one of these particles after the special suffix འུ, or the case of the following syllables found in the official lists: ནོརབུ་ རིལབུ་ ཛེབུ་ (we are not sure if these are errors or if other fusion with བུ or other syllables are possible).

4 Conclusion

We have described all the possible combinations of a Classical Tibetan syllable, listing all constraints and exceptions we found, resulting in a complete set of rules easy to implement in a computer language.

An immediate application has been to implement the rules in a spell checker running with the *hunspell* library,[23] freely available on https://github.com/eroux/hunspell-bo. Our formalization of the འུ and འ endings allowed us to implement stricter rules in our spell checker and detect more potential mistakes.

We have also managed to build solid rules for the decomposition of a Classical Tibetan syllable, listing ambiguities and possible disambiguation.

References

Bacot, Jacques. 1946. *Grammaire du Tibétain littéraire*. Paris: Librairie d'Amérique et d'Orient.

Beyer, Stephan. 1992. . Albany (NY): State University of New York Press.

[23] Hunspell (https://hunspell.github.io/) is the most popular spell checking library, used in all free software (LibreOffice, Firefox, etc.), but also in many closed source software such as the Adobe Suite, Mac OSX, etc. this makes our spell checker easily usable.

Das, Sarat Chandra. 1915. An introduction to the grammar of the Tibetan language. Darjeeling: Darjeeling Branch Press.

Dorje, Pema; and Hanfutin; and Drakpa. 2003. དག་ཡིག་གསར་བསྒྲིགས (The new spelling-dictionary). Xining: མཚོ་སྨོན་མི་རིགས་དབེ་སྐྲུན་ཁང་།.

Duff, Tony. 2000. The Illuminator Tibetan-English encyclopedia dictionary. Kathmandu: Padma Karpo Translation Committee.

Dzongkha Development Commission. 2010. ཆོས་སྐད་རྫོང་ཁ་ཚིག་མཛོད་ཆེན་མོ། (A comprehensive Chöke-Dzongkha dictionary), at http://www.dzongkha.gov.bt/en/publications/title/a-comprehensive-ch-ke-dzongkha-dictionary

Dzongkha Development Commission. 2011. བརྡ་ཡིག་གསེར་གྱི་ཤ་ལོང་། (A Handbook of Dzongkha and Chöké Abbreviations), at http://www.dzongkha.gov.bt/en/publications/title/a-handbook-of-dzongkha-and-ch-k-abbreviations

Dzongkha Development Commission. 2010. གཞི་རིམ་རྫོང་ཁའི་བརྡ་གཞུང་། (Basic Level Dzongkha Grammar Textbook), at http://www.dzongkha.gov.bt/dz/publications/title/basic-level-dzongkha-grammar-textbook

Hildt, Hélios. 2016. Towards describing Tibetan syntax: From word segmentation to rewrite rules through a semi-automatic workflow. *Himalayan Linguistics* 15: 78–112.

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

Kirtivajra, 1982. བོད་ཀྵར་གྱི་བརྡའ་ཡིག་མིན་ཚོག་དོན་གསུམ་གསལ་བྱེད། (Tibetan-Mongol dictionary, the elucidation of the three configurations of words) [Reprint in: Four Tibetan-Mongolian Lexicon. New Delhi: Sharada Rani], at https://www.tbrc.org/#!rid=W00KG09211

Negi, J.S. 1993. བོད་སྐད་དང་ལེགས་སྦྱར་གྱི་ཚིག་མཛོད་ཆེན། (Tibetan-Sanskrit dictionary). Sarnath: Central Institute of Higher Tibetan, at http://tbrc.org/link?RID=W1KG5422

Tournadre, Nicolas; and Dorje, Sangda. 2005. Manual Of Standard Tibetan: Language and civilization. Ithaca (NY): Snow Lion.

Tsering, Tashi. 2006. Standardizing Tibetan terms of information technology. The China Tibetology Research Center, at http://digitaltibetan.org/images/e/e0/Standardizing_Tibetan_Terms_of_IT-China_Tibetology_Research_Center-2006.pdf

Yisun, Zhang. 1985. བོད་རྒྱ་ཚིག་མཛོད་ཆེན་མོ། (Great Sino-Tibetan Dictionary). Beijing: མི་རིགས་དབེ་སྐྲུན་ཁང་།, at http://tbrc.org/link?RID=W29329

མཁར་སྟོང་རྫོ་རྗེ་དབང་ཕྱུག. 1979. དུས་གསུམ་རེའུ་མིག་ཐུ་མིའི་དགོངས་གཏེར། (The Treasure of Thomi's Insight, The Complete Verbal Forms). New Delhi: C.T. Khrto, at http://tbrc.org/link?RID=W1KG14490

ཚེ་ཏན་ཞབས་དྲུང་།. 2003. བོད་གངས་ཅན་གྱི་སྒྲ་རིག་པའི་བསྟན་བཅོས་ལེ་ཚན་འགའ་ཞིག་བསྡུས། (A compilation of a few themes from the grammatical treatises of the Land of Snow). Xining: མཚོ་སྔོན་མི་རིགས་དཔེ་སྐྲུན་ཁང་།, at http://tbrc.org/link?RID=W1KG25350

ཆོས་ཀྱི་འཕྲུང་གནས། (Ed.), 1721. བཀའ་འགྱུར། (Tibetan Buddhist canon), རྒྱ་དགེ། [Reproduced: Esukhia, 2012-2018], at https://github.com/Esukhia/derge-kangyur

དུང་དཀར་བློ་བཟང་འཕྲིན་ལས།, 2002. དུང་དཀར་ཚིག་མཛོད་ཆེན་མོ། (Dungkar's great dictionary). Beijing: China Tibetology Publishing House, at http://tbrc.org/link?RID=W26372

སྨོན་ལམ།. 2016. སྨོན་ལམ་ཚིག་མཛོད་ཆེན་མོ། (Monlam Grand Tibetan dictionary), at http://www.monlamit.com/

ས་ཆེན་ཀུན་དགའ་སྙིང་པོ།. 1092. ཡི་གེའི་བབྱོག་ཐབས་བྱིས་པ་བདེ་བླག་ཏུ་འཇུག་པ། (A reading manual, an accessible introduction for children) [in ས་སྐྱ་བཀའ་འབུམ། (The collected works of Sakya masters), རྒྱ་དགེ. 1736. Reprint in: Dehradun. 1992], at http://tbrc.org/link?RID=O01CT0026|O01CT00264CZ121847$W22271

ཞི་ཀྲ་པཉ་ཆེན་ཆོས་ཀྱི་འཕྲུང་གནས།, XVIIIth Century. ཡུལ་གངས་ཅན་པའི་བརྡ་ཡང་དག་པར་སྦྱོར་བའི་བསྟན་བཅོས་ཀྱི་ཟྭ་ཐག་སྒམ་ནུ་པ་དང་རྟགས་ཀྱི་འཇུག་པའི་གཞུང་གི་རྣམ་པར་བཤད་པ་མཐས་པའི་མཁྱལ་རྒྱན་རྒྱ་ཏྲིག་ཟེང་མཇོས། (The beautiful rosary of pearls ornamenting the throat of scholars, a complete presentation of the thirty root verses and the guide to signs, two treatises showing the correct application of Tibetan grammar) [Reprint in: Xining: མཚོ་སྔོན་མི་རིགས་དཔེ་སྐྲུན་ཁང་།. 2001), at http://tbrc.org/link?RID=W22787

Elie Roux
elie.roux@telecom-bretagne.eu

Hélios Hildt
hhdrupchen@gmail.com

APPENDIX: The list of valid roots and exceptions

The following pages contain the complete lists of valid roots and exceptions, easily deducible from this article. They are in a simple form:

RootOrSyllable/PropertySuffix

Where PropertySuffix is:

- A if any vowel + suffix or affixed particle can appear after the root, with the exception of the suffix ཨ
- NB if any vowel + suffix or affixed particle can appear after the root, but at least one has to appear
- C if only affixed particles can appear after the root or syllable

Note that we take special suffixes into account separately.

As an example, ཀ/A means that all the following possibilities are valid:

ཀ, ཀཀ, ཀཁས, ཀང, ཀངས, ཀད, ཀན, ཀབ, ཀབས, ཀམ, ཀམས, ཀཡ, ཀའི, ཀའིའོ, ཀའོ, ཀའང, ཀའམ, ཀར, ཀས,
ཀི, ཀིཀ, ཀིཁས, ཀིང, ཀིངས, ཀིད, ཀིན, ཀིབ, ཀིབས, ཀིམ, ཀིམས, ཀིཡ, ཀིའི, ཀིའིའོ, ཀིའོ, ཀིའང, ཀིའམ, ཀིར, ཀིས,
ཀུ, ཀུཀ, ཀུཁས, ཀུང, ཀུངས, ཀུད, ཀུན, ཀུབ, ཀུབས, ཀུམ, ཀུམས, ཀུཡ, ཀུའི, ཀུའིའོ, ཀུའོ, ཀུའང, ཀུའམ, ཀུར, ཀུས,
ཀེ, ཀེཀ, ཀེཁས, ཀེང, ཀེངས, ཀེད, ཀེན, ཀེབ, ཀེབས, ཀེམ, ཀེམས, ཀེཡ, ཀེའི, ཀེའིའོ, ཀེའོ, ཀེའང, ཀེའམ, ཀེར, ཀེས,
ཀོ, ཀོཀ, ཀོཁས, ཀོང, ཀོངས, ཀོད, ཀོན, ཀོབ, ཀོབས, ཀོམ, ཀོམས, ཀོཡ, ཀོའི, ཀོའིའོ, ཀོའོ, ཀོའང, ཀོའམ, ཀོར, ཀོས

དཀཱ/NB implies the following valid possibilities:

དཀཱཨ, དཀཱཀ, དཀཱཁས, དཀཱང, དཀཱངས, དཀཱད, དཀཱན, དཀཱབ, དཀཱབས, དཀཱམ, དཀཱམས, དཀཱཡ, དཀཱའི, དཀཱའིའོ, དཀཱའོ, དཀཱའང, དཀཱའམ, དཀཱར, དཀཱས, དཀི, དཀིཀ, དཀིཁས, དཀིང, དཀིངས, དཀིད, དཀིན, དཀིབ, དཀིབས, དཀིམ, དཀིམས, དཀིཡ, དཀིའི, དཀིའིའོ, དཀིའོ, དཀིའང, དཀིའམ, དཀིར, དཀིས, དཀུ, དཀུཀ, དཀུཁས, དཀུང, དཀུངས, དཀུད, དཀུན, དཀུབ, དཀུབས, དཀུམ, དཀུམས, དཀུཡ, དཀུའི, དཀུའིའོ, དཀུའོ, དཀུའང, དཀུའམ, དཀུར, དཀུས, དཀེ, དཀེཀ, དཀེཁས, དཀེང, དཀེངས, དཀེད, དཀེན, དཀེབ, དཀེབས, དཀེམ, དཀེམས, དཀེཡ, དཀེའི, དཀེའིའོ, དཀེའོ, དཀེའང, དཀེའམ, དཀེར, དཀེས, དཀོ, དཀོཀ, དཀོཁས, དཀོང, དཀོངས, དཀོད, དཀོན, དཀོབ, དཀོབས, དཀོམ, དཀོམས, དཀོཡ, དཀོའི, དཀོའིའོ, དཀོའོ, དཀོའང, དཀོའམ, དཀོར, དཀོས

And ཀིཌ/C implies ཀིཌ, ཀིཌའི, ཀིཌའིའོ, ཀིཌའོ, ཀིཌའང, ཀིཌའམ, ཀིཌར, ཀིཌས.

These lists constitute the basis of the spell checker we built and are available on our git repository.

We have tested our spell checker against the data of (Hildt, 2016) and have only found expected discrepancies, due to the treatment of syllables built on ཏ, ཐ, སྟ, ཋ and སྟཱ as exceptions.

Regular:

ཀ/A
ཀྱ/A
ཀྲ/A
ཀླ/A
དཀ/NB
དཀྱ/A
དཀྲ/A
བཀ/NB
བཀྱ/A
བཀྲ/A
བཀླ/A
རྐ/A
རྐྱ/A
ལྐ/A
སྐ/A
སྐྱ/A
སྐྲ/A
བརྐ/A
བརྐྱ/A
བསྐ/A
བསྐྱ/A
བསྐྲ/A
ཁ/A
ཁྱ/A
ཁྲ/A
མཁ/NB
མཁྱ/A
མཁྲ/A
འཁ/NB
འཁྱ/A
འཁྲ/A
ག/A
གྱ/A
གྲ/A
གླ/A
དག/NB
དགྱ/A
དགྲ/A
བག/NB
བགྱ/A
བགྲ/A
མག/NB
མགྱ/A
མགྲ/A
འག/NB
འགྱ/A
འགྲ/A
རྒ/A
རྒྱ/A
ལྒ/A
སྒ/A
སྒྱ/A
སྒྲ/A
བརྒ/A
བརྒྱ/A
བསྒ/A
བསྒྱ/A
བསྒྲ/A
ང/A
དང/NB
མང/NB
རྔ/A
ལྔ/A
སྔ/A
བརྔ/A
བསྔ/A
ཅ/A
གཅ/NB
བཅ/NB
ལྕ/A
ཆ/A
མཆ/NB
འཆ/NB
ཇ/A
མཇ/NB
འཇ/NB
རྗ/A
ལྗ/A
བརྗ/A
ཉ/A
གཉ/NB
མཉ/NB
རྙ/A
སྙ/A
བརྙ/A
བསྙ/A
ཏ/A
གཏ/NB
བཏ/NB
རྟ/A
ལྟ/A
སྟ/A
བརྟ/A
བལྟ/A
བསྟ/A
ཐ/A
མཐ/NB
འཐ/NB
ད/A
དྲ/A
གད/NB
བད/NB
མད/NB
འད/NB
འདྲ/A
རྡ/A
ལྡ/A
སྡ/A
བརྡ/A
བལྡ/A
བསྡ/A
ན/A
གན/NB
མན/NB
རྣ/A
སྣ/A
བརྣ/A
བསྣ/A
པ/A
པྱ/A
པྲ/A
དཔ/NB
དཔྱ/A
དཔྲ/A
ལྤ/A
སྤ/A
སྤྱ/A
སྤྲ/A
ཕ/A
ཕྱ/A
ཕྲ/A
འཕ/NB
འཕྱ/A
འཕྲ/A
བ/A
བྱ/A
བྲ/A
བླ/A
དབ/NB
དབྱ/A
དབྲ/A
འབ/NB
འབྱ/A
འབྲ/A
རྦ/A
ལྦ/A
སྦ/A
སྦྱ/A
སྦྲ/A
མ/A
མྱ/A
དམ/NB
དམྱ/A

ཀ/A
ཁ/A
ག/A
ང/A
ཅ/A
ཆ/A
ཇ/A
ཉ/A
ཏ/A
ཐ/A
ད/A
ན/A
པ/A
ཕ/A
བ/A
མ/A
ཙ/A
ཚ/A
ཛ/A
ཝ/A
ཞ/A
ཟ/A
འ/A
ཡ/A
ར/A
ལ/A
ཤ/A
ས/A
ཧ/A
ཨ/A

Exceptions:

ཀྵ/NB
ཁྱ/NB
ཁྲ/NB
གྲ/NB
གྱ/NB
ཅྲ/NB
ཆྲ/NB
ཇྲ/NB
ཉྲ/NB
ཏྲ/NB
ཐྲ/NB
དྲ/NB
དྷ/NB
དྲྷ/NB
པྲ/NB
ཕྲ/NB
བྲ/NB
བྷ/NB
བྲྷ/NB
མྲ/NB
ཙྲ/NB
ཚྲ/NB
ཛྲ/NB
ཞྲ/NB
ཟྲ/NB
ཤྲ/NB
ས་ལྷ
ས་ཡ
ས་ར
ས་ལ
ས་ཝ
ཧྲ/NB
ཧྭ/NB
ཨྱ/NB
ཨྲ/NB
ཨླ/NB
ཨྭ/NB

Specials suffix:

ཀྱི/C
ཀྱུ/C
ཀྲུ/C
ཀྲི/C
ཀླི/C
ཀྭི/C
ཁྱི/C
ཁྲི/C
གྱི/C
གྱུ/C
གྲི/C
གྲུ/C
ཅྱི/C
ཅྲི/C
ཆྱི/C
ཆྲི/C
ཇྱི/C
ཇྲི/C
ཉྱི/C
ཉྲི/C
ཏྲི/C
ཏྲུ/C
ཐྲི/C
ཐྲུ/C
དྲི/C
དྲུ/C
དྷི/C
དྷུ/C
དྲྷི/C
དྲྷུ/C
ནྱི/C
ནྲི/C
པྱི/C
པྲི/C
ཕྱི/C
ཕྲི/C
བྱི/C
བྲི/C
བྷི/C
བྷུ/C
བྲྷི/C
བྲྷུ/C
མྱི/C
མྲི/C
ཙྱི/C
ཙྲི/C
ཚྱི/C
ཚྲི/C
ཛྱི/C
ཛྲི/C
ཝྱི/C
ཞྱི/C
ཞྲི/C
ཟྱི/C
ཟྲི/C
འྱི/C
ཡྱི/C
རྱི/C
རྲི/C
ལྱི/C
ཤྱི/C
ཤྲི/C
སྱི/C
སྲི/C
ཧྱི/C
ཧྲི/C
ཧྭི/C
ཨྱི/C
ཨྲི/C
ཨླི/C
ཨྭི/C

ཨེའུ/C
སིའུ/C
སེའུ/C
སྲིའུ/C
སྲེའུ/C
བསེའུ/C
ཙའུ/C

ཀྱ/C
ཀྱི/C
ཁྱ/C
གྱ/C
གྲྭ/C
ཉ/C
ངྱ/C
དྲངས
དྲགས
ཏྲ/C
ཐྲ/C
ཕྲ/C
ཚྲ/C
ཚྭ
ཞྲ/C
ཟྲ/C
ཤྲ/C
ཧྲ/C
ཨྲ/C
བསྲ/C
བསྟྲ/C

ཀ/C
ཀཱ
ཀཱས

Wasurs: