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

Permalink
https://escholarship.org/uc/item/70z8069f

Journal
Himalayan Linguistics, 17(1)

Authors
Roux, Élie
Hildt, Hélios Drupchen

Publication Date
2018

DOI
10.5070/H917135529

License
CC BY-NC-ND 4.0

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

Elie Roux
Buddhist Digital Resource Center

Héliois Hildt
Buddhist Digital Resource Center

ABSTRACT
This document presents our research on 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
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 extent 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. We then checked these differences against some chosen Tibetan dictionaries and the Derge 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.

---

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་ȣ་པཎ་ཆེན་ཆོས་ཀྱི་འɏང་གནས།, XVIIIth c.) and (ཚǃ་ཏན་ཞབས་ȯང་།, 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), (ɥོན་ལམ།, 2016), (ȭང་དཀར་ âm་བཟང་འɋིན་ལས།, 2002), and in a lesser extent (Yisun, 1985), (ཆོས་ǰད་ɲོང་ཁ་ཚིག་མཛǑད་ཆེན་མོ།, 2010) and (Negi, 1993). We also consulted (Hill, 2010).

3 (ཆོས་ཀྱི་འɏང་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 བཇེངས་ results in radical letter ག, prefix བ, 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. 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 བྱིམ་) contain a wasur (◌ྭ). These are traditionally considered as subscribed letters, with the ability to combine with other subscribed (as in ག). 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 བྱིམ་ would be decomposed as radical letter ག, wasur, first suffix ང, second suffix ཉ and ག as radical letters, subscribed ཡ and wasur.

A second problem is the མ “suffix” (as in ཁེར་, “chapter”, /lheu/, 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 ད. For instance ཐེ་ is decomposed as radical letter ཐ and affixed particle ར. Two affixed particles can even be combined, as in ཐེ་རེ་.

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”:

\[
\text{syllable} = [\text{root}][\text{vowel}]?[\text{final part}]?
\]

\[
\text{root} = [\text{prefix}]?[\text{main-stack}]
\]

\[
\text{main-stack} = [\text{superscribed}]?[\text{radical}]?[\text{subscribed}]?[\text{wasur}]?
\]

\[
\text{final-part} = ([\text{special-suffix}][\text{affixed-particles}]?) | [\text{suffixes}] | [\text{affixed-particles}]
\]

\[
\text{affixed-particles} = [\text{affixed-particle}][\text{affixed-particle}]?
\]

For the sake of simplicity, we will not consider that the “a” vowel accent or the first suffix ད 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 multiple decomposition possibilities according to this scheme. For instance ལོ་ can be decomposed as either:

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

This schema also doesn’t account for implied suffix ཁ; for instance if བཀ་ didn’t have its affixed particle, it would be བཀ, 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 བ འ ར ལ བ འ ར or བ
- first suffix can be བ འ ར ལ བ འ ར བ འ ར or བ
- second suffix can be བ འ ར བ འ ར or བ
- superscribed can be བ འ ར or བ
- subscribed can be བ འ ར or བ
- radical letter can be བ འ ར བ འ ར བ འ ར བ འ ར or བ
- vowel accent can be བ འ ར or བ
- wasur is བ
- affixed particle can be བ འ ར བ འ ར or བ

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

- བ འ ར བ འ ར བ འ ར བ འ ར བ འ ར བ འ ར བ འ ར བ འ ར or བ for superscribed བ + radical letter
- བ འ ར བ འ ར བ འ ར བ འ ར བ འ ར བ འ ར བ འ ར བ འ ར བ འ ར བ འ ར བ འ ར བ འ ར or བ for superscribed བ + radical letter

5 Old Tibetan also has བ.
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 \textit{ཁྲིམས་མȬན་གྱི་ཚǃག་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་ཇོང་) or Dzongkha; but we chose to include 1. onomatopoeia (Ȍ་Ȍ་), 2. regional orthographies (ɺལ་Ȍ་), 3. words coming from Old Tibetan (བȰ་ȟིང་), 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 \(r\)

The subscribed \(r\) has the following possible combinations with radical letters in all the grammars we consulted: \(\text{དཀའ}, \text{གླ}, \text{གེ}, \text{ཞ}, \text{཈}, \text{ཉ}, \text{ཊ}, \text{ི}, \text{ིས}, \text{ིར}, \text{ིལ}, \text{ིོ}, \text{ིོན}, \text{ིོས}, \text{ིོྦ} \) and \(\text{ི} \). But some sources list additional possibilities: \(\text{ི}, \text{ི}, \text{ི}, \text{ི} \) and \(\text{ི} \). Among these, \(\text{ི} \) and \(\text{ི} \) are extremely rare and appear only in the following words: \(\text{ི་དཀའ} \) (“mango”) and \(\text{ི་ལྷུན} \) (O), so we consider the syllables \(\text{ི} \) and \(\text{ི} \) as exceptions\(^7\). \(\text{ི} \) is present in many words so we can count it as regular. \(\text{ི} \) is present only in \(\text{ི} \) (“ape”), \(\text{ི་བརྟེན} \) (“powder”) and \(\text{ི} \) (name of a particular

---

\(^6\) Even the very common person names \(\text{པ་ཁྲིམས་} \) and \(\text{པ་ཞེས་} \) 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 \(\text{ི་} \) would not be considered legal.
design), so we consider the syllables  and  as exceptions. We have not been able to find any word with the main stack  except proper nouns, so we don’t consider it as a possibility at all.

For combinations with radical letter and superscribed , all sources indicate  and . Some indicate  and . is present only in rare words such as (two lunar mansions), and (OT, “disordered”);  is only found in  (“speak”), (OT, “speak”) and derivatives: ,  (OT), , so we consider only these exceptions as valid for  and .

One of the sources not citing  nor  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  and . It is also possible to find  and  in some others. Our research showed that the word  (contraction of _OT) 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, so we chose to treat the syllable with wasurs as exceptions and list them. We have been able to list the following 25 syllables:

\[
\begin{align*}
\text{ཀ་} & (O) \\
\text{ཀའ} & (O) \\
\text{ཁ་} & \\
\text{ག་} & \\
\text{ʦ་} & \\
\text{ʭོ} & (OT) \\
\text{ʭངས} & \\
\text{ʭགས} & \\
\text{Ѻϫ} & \text{(OT)} \\
\text{ʏ} & \text{(OT)} \\
\text{ʌ} & \\
\text{ʴ} & \\
\text{ʴབ} & \\
\text{ʷ} & \\
\text{ʸ} & \\
\text{ʻ} & \\
\text{ʼ} & \\
\text{ʼ} & \\
\text{ˀ} & \text{ Wrath} \\
\text{ʼ} & \text{ (in ་ Wrath, OT for ཏི་, “honey”)  Wrath.}
\end{align*}
\]

2.2.5 Constraints on prefix+radical

The following rules on prefixes are found in all grammars:

-  can only be prefix of  or  or  or  or
-  of  or  or
-  of  or  or
-  of  or  or
-  of  or  or

But (Tournadre and Dorje, 2005) gives other possibilities for prefix  combined with  and . This difference can easily be explained: in these cases,  cannot be prefix of the radical letter when there is no superscribed (ex: འིང་ is impossible), it only can when a superscribed is present (ex: འིང་).
2.2.6 Constraints on root

The main additional constraint found in traditional grammar is that prefixes ཕ and བ cannot appear when a superscribed letter is present.

But (ཅྱི་ཧོ་ཁང་ཐོ་, XVIIIth c.) goes in more details and gives a more specific list of possible combinations of roots. Leaving aside the wasurs and the naked radical letters, the given possibilities are the following:

(ཚོན་མཁན་ཤོར, 2003) also gives two exceptions of syllables that do not follow these constraints: བགླ་ and མȮོན་. We were able to find མȮོན་ as an abbreviation of མདོ་Ȯན་ ("to remember a sutra") in (བོད་ཡིག་གསེར་གྱི་ཨ་ལོང་ / A Handbook of Abbreviations, 2011), but did not find བགླ་ in the sources we consulted, so we do not consider them valid. We also found བཙོ་ cited in (Kirtivajra, 1982), but we are not sure of its relevance.

As noted earlier, our proposal is to also to remove ȁ, Ȅ, Ȇ, Ȯ, Ɇ and ɋ 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:

• བརྗོ་ found only in བརྗོང་ and བརྗོངས (both: "to extend")
• བཀླ found only in བཀླག་, བཀླགས་ (both: "to read out loud"), བྲན་ and བྲན་ (both: "to wear")
• བྲི found only in བྲི་ ("to smile") and བྲི་ ("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 མ can only be found after the first suffixes བ or མ.

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

For first suffix བ, the rule given in (Beyer, 1992) can be easily verified; we choose to formalize it as follows: first suffix བ 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 བ:

11 Although the list is not given directly, it is quite straightforward to infer, and it is given more directly in (Das, 1915) and (ཚོན་མཁན་ཤོར, 2003).
12 Although they have been found in (ཚོན་མཁན་ཤོར, 1092), which we did not have time to fully explore for this article.
The only legal syllables with suffix \(\alpha\) not in this list are \(ནེའོ\) (OT, AS for \(ནེ\), “unclean”) and \(བȦའ\) (OT, AS for \(བȦ\), “sign”).13 We can also find the onomatopoeia \(མེའ\) (O), in which the suffix \(\alpha\) cannot be replaced by an affixed particle, which could maybe be pushed forward as in the (extremely unlikely) \(མེའའི\).

### 2.2.8 Constraints on affixed particles

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

The pronunciation of affixed particles and special suffixes 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 \(\beta\):

\[
\begin{align*}
\text{མིའོ} & \quad \text{ཐོམེའོ} & \quad \text{ཨེའོ} & \quad \text{ཨོ} & \quad \text{ཨེཆེའོ} & \quad \text{ཨོི} & \quad \text{ཨོོ} & \quad \text{ཨོོོ} \\
\text{ཨོོོ} & \quad \text{ཨོོོ} & \quad \text{ཨོོོ} & \quad \text{ཨོོོ} & \quad \text{ཨོོོ} & \quad \text{ཨོོོ} & \quad \text{ཨོོོ} & \quad \text{ཨོོོ}
\end{align*}
\]

(OT)\(,\)  \(\text{AS for}\) \(ཨོ\) (found in Chinese words), \(ཨོ\) \(\text{AS for}\) \(ཨོ\) (found in Sanskrit words), \(ཨོ\) (Mongolian), \(ཨོ\) \(\text{AS for}\) \(ཨོ\) (found in proper names).

Other special suffixes appear exceptionally, they are \(\alpha\) in \(ནེ\) (R, AS for \(ནེ\), “buckwheat”), \(ནེ\) (OT, “conversation”) and \(གོ\) (O).16

---

13 We did not consider \(བȦའ\) given in (མཁར་Ȫོད་Ȱོ་Țེ་དབང་Ɋག, 1979) as this source is not considered reliable.

14 Note that this allows to build written syllables with the 4 vowels such as \(ལེར\) ("the one of the chapter"), pronounced \(\text{[lewy}:\text{jo}:\}\) (3 syllables) by one informant with a Standard Tibetan accent.

15 It doesn't include \(\text{ɡɹ}\) (found in Chinese words), \(\text{ɡɹ}\) (found in Sanskrit words), \(\text{ɡɹ}\) (Mongolian), \(\text{ɡɹ}\) \(\text{AS for}\) \(ɡɹ\) (found in proper names).

16 (Tourmadre and Dorje, 2005) also cites \(\alpha\) 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 བའི་, ལེ་, བའམ་, etc. Based on the information we gathered from our research, we propose to apply the following rules in this particular order:

1. if the syllable contains a subscript, superscript or wasur then the main stack is what contains it (ex: རྙེད་ → ར)
2. if a letter other than འ carries a vowel then it is the main stack (གཞི་ → ཐ)
3. if a vowel is carried by འ and འ is the first letter then འ is the main stack (འོད་ → འ)
4. if a vowel accent is carried by an འ which is not the first letter, the main stack is before the first འ with a vowel

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 གང or གམ, then the main stack is right before གང or གམ
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 ལ, then the main stack is the second letter
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
11. if ལ cannot be second suffix after the second letter if it was a first suffix (meaning the second letter is not ལ, ལ, ལ nor ལ), then the main stack is the second letter

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 ལ, 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 ས. The only exception to this rule is the very unlikely ཚོ་ཤ (O), but the legality of this syllable is not clear.
19 This rule could be extended for Old Tibetan with ཅ. 

57
is $g$, $c$, $m$ or $s$. It is easy to find the 9 corresponding cases: $མངས$, $མག$, $དབ$, $དང$, $དག$, $དམ$, $བག$, $འབ$, and $འག$. For each of these cases we have three possible structures:

1. $s$ is a second suffix (e.g. $མངས$ can be decomposed as radical $m$, suffix $ང$, second suffix $s$, we note this decomposition $m|ངs$)
2. $s$ is a suffix (e.g. $མང$ is radical $ང$, prefix $m$, suffix $s$, noted $མང|s$)
3. $s$ is an affixed particle (e.g. $མང$ radical $ང$, prefix $m$, suffix $ལ$ replaced by affixed particle $s$, noted $m|ལs$)

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:

- $མང$: $m|ցs$, $m|ցs$, $m|ցs+a+g$ → ambiguous, $m$ as main stack is more intuitive to our Tibetan informants
- $འབ$: *$a|ցm$, *$a|ցm$, *$a|ցm+a$ → $a$ is the main stack
- $མག$: *$s|ցm$, *$m|ցm$, *$m|ցm+a$ → ambiguous, $m$ as main stack is more intuitive to our Tibetan informants
- $བག$: *$b|ցm$, *$b|ցm$, *$b|ցm+a+g$ → $b$
- $མ$: *$a|ցm$, *$a|ցm$, *$a|ցm+a$ → $a$
- *$t|ցm$, *$t|Ցm$, *$t|Ցm+a+g$ → $t$
- *$t|ցm$, *$t|Ցm$, *$t|Ցm+a+g$ → $t$
- *$t|ցm$, *$t|Ցm$, *$t|Ցm+a+g$ → $t$
- *$t|ցm$, *$t|Ցm$, *$t|Ցm+a+g$ → $t$
- $ད$: *$g|Ցm$, *$g|Ցm$, *$g|Ցm+a$ → $g$
- *$g|Ցm$, *$g|Ցm$, *$g|Ցm+a$ → $g$
- *$g|Ցm$, *$g|Ցm$, *$g|Ցm+a$ → $g$
- *$g|Ցm$, *$g|Ցm$, *$g|Ցm+a$ → $g$
- *$g|Ցm$, *$g|Ցm$, *$g|Ցm+a$ → $g$
- *$g|Ցm$, *$g|Ցm$, *$g|Ցm+a$ → $g$

### 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 $s$ and $x$, plus the syllables $m$, $m$, $m$. 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 སི, ས can only be affixed particle, because radical ས + 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 སར, 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 on 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 (ཚǃ་ཏན་ཞབས་ぉང།, 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་ɲོད་ཁའི་བȰ་ག鸥ང།, 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 ɹ forms

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

\[
\begin{align*}
\text{ཅིɹ} & \quad \text{ཇིɹ} \\
\text{ཏɹ} & \quad \text{དɹ} \\
\text{ནིɹ} & \quad \text{བོɹ} \\
\text{ཙɹ} & \quad \text{ཚིɹ} \\
\text{ཞིɹ} & \quad \text{ȟིɹ} \\
\text{Ȧིɹ} & \quad \text{Ȩེɹ} \\
\text{ʈིɹ} & \quad \text{ཤིɹ} \\
\text{ȴིɹ} & \quad \text{ཧɹ} \\
\text{ཧིɹ} & \quad \text{གྭɹ} \\
\text{ʭɹ} & \quad \text{ʸɹ} \\
\end{align*}
\]

#### 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: ག ག ག ག. All these particles are written at the end of a syllable in their usual form, except ག which is noted ག (as in ག, not ག, which would introduce some ambiguity with the ག suffix). We were able to infer the following rules:

- no other affixed particle can be affixed after these particles (ex: ག is not correct)
- unlike other affixed particles, they do not replace the ག suffix (ex: ག is correct)
- possible combinations of suffix + particles are given in (གཞི་རིམ་ɲོད་ཁའི་བྱཀྲ་ག鸥ང།, 2010), we reproduce the table here:

---


\(^{22}\) Available on https://github.com/eroux/hunspell-dz.
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 ɹ, or the case of the following syllables found in the official lists: ནོ་ རི་ ལི (we are not sure if these are errors or if other fusion with ནོ 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 ɹ and ཡ 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


---

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.

Dorje, Pema; and Hanfutin; and Drakpa. 2003. བྱུགས་ལོངས་དབུ་བྱུགས་བོས་(The new spelling-dictionary). Xining: ཀིགས་ཤིག་རིགས་དཔེ་གན་ཁང་།.


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

Kirtivajra, 1982. ཀིགས་རིམ་བོད་གྱི་བོད་དཔེ་གན་ཁང་། (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


Roux and Hildt: Algorithmic description of the decomposition of a Classical Tibetan syllable

Elie Roux
elie.roux@telecom-bretagne.eu

Hélios Hildt
hhdrupchen@gmail.com

62
**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 α
- 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, α/A means that all the following possibilities are valid:

As exceptions, α/NB implies all the following possibilities:

And α/C implies α as exceptions.

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 ง, 琰, ฎ and ฏ as exceptions.
<table>
<thead>
<tr>
<th>Regular:</th>
<th>᥄/NB</th>
<th>᥄/NB</th>
<th>᥄/A</th>
<th>᥄/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>᥄/NB</td>
<td>᥄/A</td>
<td>᥄/A</td>
<td>᥄/A</td>
<td>᥄/A</td>
</tr>
<tr>
<td>᥄/A</td>
<td>᥄/A</td>
<td>᥄/A</td>
<td>᥄/A</td>
<td>᥄/A</td>
</tr>
<tr>
<td>᥄/A</td>
<td>᥄/NB</td>
<td>᥄/NB</td>
<td>᥄/A</td>
<td>᥄/A</td>
</tr>
<tr>
<td>᥄/A</td>
<td>᥄/A</td>
<td>᥄/A</td>
<td>᥄/A</td>
<td>᥄/A</td>
</tr>
<tr>
<td>᥄/A</td>
<td>᥄/A</td>
<td>᥄/A</td>
<td>᥄/A</td>
<td>᥄/A</td>
</tr>
<tr>
<td>᥄/A</td>
<td>᥄/A</td>
<td>᥄/A</td>
<td>᥄/A</td>
<td>᥄/A</td>
</tr>
</tbody>
</table>

Roux and Hildt: Algorithmic description of the decomposition of a Classical Tibetan syllable
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>ɣ</td>
<td>ɣ/NC</td>
</tr>
</tbody>
</table>

**Specials suffix:**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ɣ</td>
<td>ɣ/C</td>
</tr>
</tbody>
</table>

**Exceptions:**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ལེ་</td>
<td>ལེ་/C</td>
</tr>
</tbody>
</table>

**Specials suffix:**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ɣ</td>
<td>ɣ/C</td>
</tr>
</tbody>
</table>

65
Roux and Hildt: Algorithmic description of the decomposition of a Classical Tibetan syllable

\[
\begin{array}{ll}
\text{ཧེ}/C & \text{ི}/C \\
\text{སི}/C & \text{ི}/C \\
\text{སེ}/C & \text{ི}/C \\
\text{རི}/C & \text{ི}/C \\
\text{ʃེ}/C & \text{ི}/C \\
\text{བས}/C & \text{ི}/C \\
\text{ཏ}/C & \text{ི}/C \\
\text{Wasurs:} & \\
\text{ཀྭ}/C & \\
\text{ཀྭའི}/C & \\
\text{ཁྭ}/C & \\
\text{གྭ}/C & \\
\text{ʋ}/C & \\
\text{ʦ}/C & \\
\text{ʭོ}/C & \\
\text{ʭངས}/C & \\
\text{ʭོ}/C & \\
\text{ʴ}/C & \\
\text{ʴབ}/C & \\
\text{ʷ}/C & \\
\text{ʸ}/C & \\
\text{ʻ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\text{ʼ}/C & \\
\end{array}
\]