UC Santa Barbara

Himalayan Linguistics

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

Copyright Information

Copyright 2018 by the author(s). This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at https://creativecommons.org/licenses/by-nc-nd/4.0/

Peer reviewed



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 © 2018. All rights reserved.

This Portable Document Format (PDF) file may not be altered in any way.

Tables of contents, abstracts, and submission guidelines are available at escholarship.org/uc/himalayanlinguistics

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 language¹. We then checked these differences against some chosen Tibetan dictionaries² and the Derge Kangyur³.

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 (Argure and England, XVIIIth c.) and (Ergare and Company). 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), (ब्रॅब्य्यज्ञा, 2016), (दुर्द्यपुत्र ब्रॅब्युव्यज्ञा, 2002), and in a lesser extent (Yisun, 1985), (क्र्युव्यक्त क्रियुव्यक्त क्रियुव्यक्त

^{3 (}রুম্প্রান্ত্র প্রমা, 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

2.1.2 Missing elements

First, some syllables (such as 5^{max}) contain a wasur (2). These are traditionally considered as subscribed letters, with the ability to combine with other subscribed (as in $\frac{1}{2}$). 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 5^{max} would be decomposed as radical letter $\frac{1}{2}$, wasur, first suffix $\frac{1}{2}$, second suffix $\frac{1}{2}$; and $\frac{1}{2}$ as radical letter, subscribed $\frac{1}{2}$ and wasur.

A second problem is the 3 "suffix" (as in ag, "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 a. For instance sale is decomposed as radical letter st and affixed particle a. Two affixed particles can even be combined, as in salat.

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 α 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 star can be decomposed as either:

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

This schema also doesn't account for implied suffix α : for instance if β_{α} did not have its affixed particle, it would be β_{α} , 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 n 5 n a or a
- first suffix can be ๆ ธ ร ส จ ส ส ส or พ
- second suffix can be \$15
- superscribed can be z a or a
- subscribed can be ୁ ୁ or ୍ମ
- radical letter can be न्वन्दकहनुन्न्द्र्यप्रवर्षस्व्वत्य्र्वन्त्र्
- vowel accent can be ି ୍ର or ි
- wasur is 4
- affixed particle can be a a as as a or a

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

- ក្រុក្ខភ្នុង 🕫 for superscribed 🥆 + radical letter
- স্ন্র্র্র্র্র for superscribed a + radical letter

. ,

⁵ Old Tibetan also has 5.

• អូតូតូតូតូ for superscribed ។ + radical letter + subscribed ។

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 handles. While there is no clear cut definition of this concept, we exclude syllables appearing solely in 1. transliterated foreign languages, 2. proper nouns⁶, 3. misspellings, 4. words in other languages such as Zhang zhung language (argaing) or Dzongkha; but we chose to include 1. onomatopoeia (handles), 2. regional orthographies (handles), 3. words coming from Old Tibetan (argaing), 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 π

⁶ Even the very common person names $\eta_{\overline{A}}$ and $\varsigma_{\overline{A}}$ 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 ₹₹ would not be considered legal.

design), so we consider the syllables \mathfrak{H} and \mathfrak{H} as exceptions. We have not been able to find any word with the main stack \mathfrak{g} except proper nouns, so we don't consider it as a possibility at all.

For combinations with radical letter and superscribed on, all sources indicate of one one indicate one of one indicate of one one indicate one of one one indicate one of one one indicate one one one one one one one one of one of one one one one of one one one of one of one one one of one of one of one of one of one one of one

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

2.2.4 Constraints on the wasur

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 words⁹, so we chose to treat the syllable with wasurs as exceptions and list them. We have been able to list the following 25 syllables:

2.2.5 Constraints on prefix+radical

The following rules on prefixes are found in all grammars:

- ব্can only be prefix of ভ নু দু বু র র নু নু নু বু বি
- **5** of শ্বাহ্র্ব or অ
- प्राचित्र विश्व विष्य विश्व विष्य विश्व विश्य
- वा of विष्टळहलुबर्वळं or हं
- ব of বিশ্ভর্রর্থন হ or র্

0

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 $\sqrt[4]{}$, $\sqrt[4]{}$ (in proper nouns), $\sqrt[4]{}$, $\sqrt[4]{}$ (Zhang zhung language), $\sqrt[4]{}$, $\sqrt[4]{}$, $\sqrt[4]{}$, $\sqrt[4]{}$ (Chinese), $\sqrt[4]{}$ (given only in (ልሚተ፞፞ዿ፟ጚ ද ድርፍሮኒ (1979)), $\sqrt[4]{}$, $\sqrt[4]{}$ nor $\sqrt[4]{}$ (Dzongkha).

2.2.6 Constraints on root

The main additional constraint found in traditional grammar is that prefixes $\P \subseteq \P$ and Q cannot appear when a superscribed letter is present.

But (মণ্ড্ৰ'ড্ৰেইড্ৰেইড্ৰ'ট্ৰ'ব্ৰেড্ৰ্ৰ্ড্), 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:

(শ্রুল্বিস্পেন্ত্র্ন), 2003) also gives two exceptions of syllables that do not follow these constraints: বন্ধ্রা and কর্ত্র্বু: We were able to find কর্ত্রু as an abbreviation of কর্ত্বুলু ("to remember a sutra") in (ব্ছুল্পিল্পিন্ট্রুল্পেন্ত্রুল্পেন্ত্রুল্পেন্ত্রুল্পেন্ত্রুল্পেন্ত্রুল্পেন্ত্রুল্পেন্ত্রুল্পেন্ত্রুল্পেন্ত্রেল্পেন্ত্রুল্পেন্ত্রেল্পেন্ত্রুল্পেন্ত্রুল্পেন্ত্রুল্পেন্ত্রুল্পেন্ত্র্ন্ত্রুল্পেন্ত্রেল্পেন্ত্রুল্পেন্ত্রেল্পেন্ত্রুল্পেন্ত্রেল্পেন্ত্রেল্পেন্ত্রেল্পেন্ত্র্ন্ত্রেল্পেন্ত্র্ন্ত্রুল্পেন্ত্রেল্পেন্ত্র্ন্ত্রেল্পেন্ত্র্ন্ত্রেল্পেন্ত্র্ন্ত্রেল্পেন্ত্র্ন্ত্র্ন্ত্রল্পেন্ত্র্ন্ত্রেল্পেন্ত্রেল্পেন্ত্র্ন্ত্রেল্পেন্ত্র্ন্ত্রেল্পেন্ত্র্ন্ত্রেল্পেন্ত্র্ন্ত্

As noted earlier, our proposal is to also to remove 5, 8, 8, and 8 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:

- স্কু found only in স্কুদ and স্কুদ্ (both: "to extend")
- ¬¬¬ found only in ¬¬¬¬ ("to smile") and ¬¬¬¬¬ ("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 n can only be found after the first suffixes $n \leq n$ or n.

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

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

¹¹ Although the list is not given directly, it is quite straightforward to infer, and it is given more directly in (Das, 1915) and (ﷺ, 2003).

¹² Although they have been found in (अ केन्द्र शुन्न पुन्न केन्द्र, 1092), which we did not have time to fully explore for this article.

वालय, वर्षिय, वर्षिय, वर्षिय, वर्षिय, वर्षिय, अर्थय, ट्राय, यस्य, ट्राय, यय्य, र्याष्ट्र, वर्ष्य, अर्थ्य, अर्थ्य, यर्ष्य, वर्ष्य, वर्य, वर्य, वर्य, वर्ष्य, व

The only legal syllables with suffix α not in this list are $\sqrt[3]{\alpha}$ (OT, AS for $\sqrt[3]{\alpha}$, "unclean") and $\sqrt[3]{\alpha}$ (OT, AS for $\sqrt[3]{\alpha}$, "sign")13. We can also find the onomatopoeia $\sqrt[3]{\alpha}$ (O), in which the suffix α cannot be replaced by an affixed particle, which could maybe be pushed forward as in the (extremely unlikely) $\sqrt[3]{\alpha}$.

2.2.8 Constraints on affixed particles

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 3:

Other special suffixes appear exceptionally, they are $\breve{\alpha}$ in $\breve{\beta}\breve{\alpha}$ (R, AS for $\breve{\gamma}\breve{\alpha}$, "buckwheat"), $\breve{\beta}\breve{\alpha}$ (OT, "conversation") and $\breve{\gamma}\ddot{\alpha}$ (O). ¹⁶

¹³ We did not consider $\sqrt{5}$ given in ($\sqrt{5}$ $\sqrt{$

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

⁽Tournadre and Dorje, 2005) also cites à 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 5th, Ag, 5th, Ag, 5th, etc. Based on the information we gathered from our research, we propose to apply the following rules in this particular order¹⁷:

- 1. if the syllable contains a subscript, superscript or wasur then the main stack is what contains it (ex: $5858^{\circ} \rightarrow 5$)
- 2. if a letter other than q carries a vowel then it is the main stack ($q | q \rightarrow q$)
- 3. if a vowel is carried by a and a is the first letter then a is the main stack ($\tilde{a}_{5} \rightarrow a$)

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 q = 0 or q = 0, then the main stack is right before q = 0 or q = 0 (q = 0)
- 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 $(\neg \neg \neg \neg)$
- 8. if the syllable contains four letters, the main stack is the second letter $(\neg \exists \neg \neg \neg \exists)$

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

- 9. if the final letter is not \sqrt{n} , then the main stack is the second letter (\sqrt{n})
- 11. if \P cannot be second suffix after the second letter if it was a first suffix (meaning the second letter is not \P , \neg , \neg nor \neg , then the main stack is the second letter (\P \P \P \P \P \P

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 51, 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 \hat{a}_N . The only exception to this rule is the very unlikely $\hat{a}_N \hat{a}_N$ (O), but the legality of this syllable is not clear.

¹⁹ This rule could be extended for Old Tibetan with 93.

is of, c, or a. It is easy to find the 9 corresponding cases: alay, and a solar, and a solar.

For each of these cases we have three possible structures:

- 2. N is a suffix (e.g. ALAY is radical 5, prefix A, suffix N, noted AL N)

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:

- sken: a| cn, *ac|n, ac|a+n → ambiguous, a as main stack is more intuitive to our Tibetan informants
- q¬¬¬; *a |¬¬¬, *a¬ |¬¬, a¬ | a+¬¬ ¬ is the main stack
- क्याका: *ब्र|पाव, *ब्राप|प्प, *ब्र|प्प, *ब्राप|प्प, *ब्राप|प्प, *ब्राप|प्प, *ब्राप|प्प, *ब्राप|प्प,
- বেশ্বম: *ব ব্লম, বেল অ, বেল ব+ম → শ্ব
- **598**: *5|58,55|8 (OT),55|6+8 → 5
- **দ্বাম**: *দ্বাম, দ্বাম, দ্বাম, দ্বাম + ম
- **5EN**: $5 \mid \pi \mid$ (misspelling of $5 \pi \mid \pi \mid \pi \mid$), * $5 \pi \mid \pi \mid \pi \mid \pi \mid \pi \mid \pi \mid$
- รุงพา: *รุ|มพ, รุม|พ, รุม| ผ+พ → ม

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".

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

The disambiguation of \P and \P can sometimes be done at syllable level according to the existence or absence of one version of the syllable; ex: in \P can only be affixed particle, because radical \P + suffix \P 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 $q_{\overline{q}}$, in which we cannot decide if q 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 software²¹ and tested it against all the data produced for (Hildt, 2016), based on (** '5 3' (** 55'), 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 (प्रविद्धार पुराविष्ट पुराविष्ट , 2010) and private discussions with members of the Dzongkha Development Commission (hereafter referenced as DDC) about a Dzongkha syllable spell checker for the *hunspell* library²². 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 sforms

The word list we received contains the following wasurs and ${}_{\mbox{\scriptsize G}}$ forms that have not been encountered in Classical Tibetan sources:

3.4.2 Additional affixed particles

- no other affixed particle can be affixed after these particles (ex: ক্রুইন্রে is not correct)
- unlike other affixed particles, they do not replace the a suffix (ex: محرمة is correct)
- possible combinations of suffix + particles are given in (ন্ত্ৰিইফাইন্ন্ত্রেন্, 2010), we reproduce the table here:

_

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

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

| | ঘ | ð | 口角 | Ť | व/ वॅ |
|-----|---|---|----|---|-------|
| ব্য | X | X | | | X |
| _ | X | X | | | X |
| 5 | X | X | X | | X |
| ब् | X | X | | | X |
| 7 | X | | | | X |
| δl | X | X | | | |
| ď | | | X | X | X |
| ㅈ | X | X | X | | X |
| ਕ | X | X | X | | X |
| Ø | 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 g, or the case of the following syllables found in the official lists: $\sqrt[3]{3} - \sqrt[3]{2} - \sqrt[3$

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 $\mathfrak g$ and $\mathfrak q$ 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. নুশ্ অল্বাল্যমন্ত্রীশ্র (The new spelling-dictionary). Xining: কার্ক্টাপ্রক্রিক্টার্ক্টার ক্ট্রেল্টার্ক্
- Duff, Tony. 2000. The Illuminator Tibetan-English encyclopedia dictionary. Kathmandu: Padma Karpo Translation Committee.
- Dzongkha Development Commission. 2010. 黃宗宗黃宗承宗 (A comprehensive Chöke-Dzongkha dictionary), at http://www.dzongkha.gov.bt/en/publications/title/a-comprehensive-ch-ke-dzongkha-dictionary
- Dzongkha Development Commission. 2011. ব্ৰুপ্ৰাণ্ডিৰ শ্ৰীজেৰ্বিল (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. 呵奇 宋京京帝 (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. জন্ত্ৰ জ্বাজ্য ক্ৰাজ্য ক্ৰাজ্য ক্ৰাজ্য ক্ৰাজ্য ক্ৰাজ্য বিশ্ব ক্ৰাজ্য ক্ৰাজ্য (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. ব্ৰুম্ব্ৰেজ্ব ক্ৰিন্ত ক্ৰিন্ত কৰিব কৰিব বিষয়ে (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

- ঝানম শূর্ন ই ই দ্বান প্রেল। 1979. দুঝানারের ঝানারের ঝানারের ঝানারের ঝানারের ঝানারের ঝানারের ঝানারের ঝানারের বাদিনারের (The Treasure of Thomi's Insight, The Complete Verbal Forms). New Delhi: C.T. Khrto, at http://tbrc.org/link?RID=W1KG14490
- ಹান্ত্ৰ ব্ৰহণ হুলা. 2003. হাল্লাল্ডৰ ট্ৰান্ত্ৰ বিষ্ণাল্ডৰ বিষ্ণালড়ৰ বিষ্ণালড বিষ্ণালড়ৰ বিষ্ণালড়ৰ বিষ্ণালড়ৰ বিষ্ণালড়ৰ বিষ্ণালড়ৰ বিষ্ণালড়ৰ বিষ্ণালড়ৰ বিষ্ণালড
- ৰ্ম শ্রীনেজন (Ed.), 1721. নান্ননেজন (Tibetan Buddhist canon), ইন্ট্রা [Reproduced: Esukhia, 2012-2018], at https://github.com/Esukhia/derge-kangyur
- নুদ্দ্ৰ্যান স্থান্ত্ৰ ব্ৰহ্ম ক্ষ্ম (Dungkar's great dictionary). Beijing: China Tibetology Publishing House, at http://tbrc.org/link?RID=W26372
- ৰ্মুৰ্'অমা. 2016. ৰ্মুৰ্'অম'স্কৰ্জ্বা মাৰ্ছ্ৰ্ ক্লিৰ্'মা (Monlam Grand Tibetan dictionary), at http://www.monlamit.com/
- শাক্তির শূর্য দ্বাদ স্থিন হাঁ. 1092. ঐপেট নিয়াল লেকা নিয়াল নিয
- बेन्द्रम्भान्ने क्रियो त्रित्त प्रमानि Century. ખુબ जून कर राते यह प्रमानि प्रमानि प्रमानि प्रमानि प्रमानि प्रमानि प्रमानि कर्म क्रिया प्रमानि प्रमा

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 a
- 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, 77/A means that all the following possibilities are valid:

ग्न, ग्रावा, ग्रावाय, ग्राव्य, ग्राव्

571/NB implies the following valid possibilities:

حسام، حسام، حسام، حساح، حساط، حساط، حساط، حساط، حساح، حساح،

And मेल्न/C implies मेल्न, मेल्लवे, मेल्लवेर, मेल्लवेर,

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 5, 8, 8 and 8 as exceptions.

| Regular: | ag/A | out/NB | _₹ /A | 55/A |
|-------------------------|--|-------------------|----------------------------|---------------------------------------|
| | ag/A | ξ/A | ۶۶/A | 55/A |
| η/A | ₽/A | ₽/A | ₽₽/A | ≅/A |
| IJ/A | g/A | _₹ /A | 75/A | _₹ /A |
| ıر/A | $_{\mathfrak{T}}\!/A$ | se/A | ₽/A | ₃ /A |
| д/A | $_{\Xi}\!/A$ | ne/A | MR/NB | §/A |
| | 59/NB | ⊗/A | $_{\text{GS}}/\mathrm{NB}$ | 54/A |
| 57J/A | 59/A | $\sim NB$ | 5/A | s/A |
| <u> দুশ্</u> য/A | 5 | ¬ᢐ/NB | 5/A | 5/A |
| na /NB | 777/NB | ₃/A | ۹۲/NB | ard/NB |
| ¬IJ/A | ag/A | ж/ А | ¬¬/NB | as/ A |
| ₽ŢĮ/A | and A | \sim | ait/NB | aş $/A$ |
| ¬д/A | and/NB | дъ/NB | a5/NB | 5/A |
| π/A | ayy $/\!A$ | E/A | a $5/A$ | s/A |
| ⊕/A | ম্মৃ $/\!A$ | $\sim NB$ | z /A | ≅/A |
| ∄⁄A | ay/NB | α E/ NB | ĕ/A | ≅/A |
| _₹ /A | ag/ A | ĕ/A | ≅/A | 5√NB |
| _Ŋ /A | $lpha_{ar{z}}/A$ | ₽/A | ۶۴/A | 55/A |
| _₹ /A | $_{\bar{\eta}}/\mathrm{A}$ | ς∉/A | ₽₽/A | 5 <u>5</u> /A |
| $\sim 10^{10} / { m A}$ | ₅ /A | ₃ /A | ন <u>হ</u> /A | an/NB |
| ¬₼/A | $_{\mathfrak{A}}/\mathbf{A}$ | শ্ৰ/NB | a/A | as/ A |
| $\sim A$ | $_{\mathbb{F}}/A$ | NB | শ্ৰ/NB | as/ A |
| and/A | $_{\S}/\mathrm{A}$ | _₹ /A | মন্/ ${ m NB}$ | 5/A |
| THA | _{ञ्} /A | _₹ /A | ₹/A | ≅/A |
| $_{	au}/A$ | ~ 4 | 勺養/A | $_{\S}/\mathrm{A}$ | ≅/A |
| g/A | ~ 5 | 53/A | ~ 4 | _Ŋ /A |
| g/A | ~ 4 | ₅ /A | ¬₹⁄A | _Ŋ /A |
| aya/NB | asj/A | य <u>ा</u> त्र/NB | 51/A | $\sigma_{\!\scriptscriptstyle 1}\!/A$ |
| $a_{\rm NS}/A$ | \\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\ | 75/NB | s/A | ₹/A |
| ang/A | =/A | ₅ /A | 5√A | 5₫/NB |
| an/NB | 55/NB | ą/A | √√NB | 53/A |
| | | | | |

| $_{\bar{\epsilon}_{\!$ | π/A | हुत् इत्य | Èg∕C | 5g/C |
|--|--------------------|-------------------------|------------------|-------------------|
| _₹ /A | a/A | 8 | 9g/C | gg/C |
| _Ŋ /A | ন _{র্} /A | Specials suffix: | ह्रे द /C | ag/C |
| _§ /A | a/A | • | नेख/C | SG/C |
| ĕ/ A | A/A | ₹ă/C | गहेद/С | ag/C |
| শ্ৰুৰ্ /NB | MA/NB | ajă/C | ₹G/C | ĕg∕C |
| বৰ্জ/NB | ₽₽/NB | MG/C | हेख/C | ন্দ্রনূ/C |
| ₹ /A | /A | मेख/С | हे द /C | gg/C |
| _₹ /A | _≅ /A | गेद्ध/С | ag/C | बेतु/С |
| As/A | a/A | तितु/C | ब्रेदु/С | 3 €/C |
| ~ 10 | apa/NB | _{बिलु} /C | बेदु/С | ag/C |
| _ぜ /A | 55/NB | BG/C | ág/C | ig/C |
| atal/NB | as/A | BG/C | मधेतु/С | ₹તુ/C |
| त्रर्ज/NB | ‹/A | ag/C | ₹g/C | ग्रहेतु/С |
| €/A | 5/A | AG/C | ₹ 6/C | €g/C |
| ME/NB | ₅ /A | ag/C | arg/C | ₹g/C |
| ae/NB | ₽/A | मुदु/С | ₹g/C | _{etg} /C |
| ≝/A | «V/A | चेद् / C | <u> हेतु</u> /C | ĕg∕C |
| η≝/A | | ag/C | इंदु/С | मळेंदु/С |
| $_{\mathfrak{A}}\!/A$ | Exceptions: | त्यो <mark>त् /C</mark> | ब्द्/С | मळेंदु/С |
| A/A | | ag/C | बेलु/С | €g/C |
| \sqrt{NB} | म्ब्रेय ब्रेय | ag/C | इें <u>द</u> /C | _{Èg} /C |
| NB | THE THE | ag/C | न्चेख∕C | aleg/C |
| ∌/A | 5/C | gg/C | ag/C | àg/C |
| ã/A | ন্ত্ৰৰ প্ৰব্য | ag/C | ₹G/C | યોતુ/C |
| ₽₽₽/NB | ĕl∕C | ₹g/C | gg/C | ग्यितु/С |
| ¬¬NB | §/C | इंदु/С | gg/C | ₹ु/С |
| 勺煮/A | A Za Za | _{चेदु} /C | sg/C | વાતુ/C |
| a/A | ed/C | य्डितु/С | झेख्/C | बेदु/С |
| w/A | <u> </u> | याडेतु/С | दचेतु/С | बेतु/С |
| ημ/NB | EL SEL | _{ਵੇਰ} /C | नेदु/С | AG/C |

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

구년/C 첫/C

최년/C 첫메

최년/C

최년/C

최년/C

최년/C

조선/C

조선/C

Wasurs:

77/C गुत्रे/С ₁₃/C 47/C **A/C** 3/C <u>z</u>/C र्नुवास र्नुवास 5/C ş/C **₹/C** <u>ಹ</u>್/C థ్ఛ⊓ α / C \sqrt{C} \sqrt{C} Ŋ/C ą/C

§/C

বন্
/C
বন্
/C