himalayan linguistics

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

Himalayan Linguistics

Towards describing Tibetan syntax: From word segmentation to rewrite rules through a semi-automated workflow

Hélios Hildt

Université Bordeaux Montaigne

ABSTRACT

The first task in Tibetan Natural Language Processing is word segmentation. We present our lightweight segmentation tool that is based on lexical resources. It can be executed within InDesign and the user can update it with the manual corrections of its output. We then propose a semi-automated workflow aiming at syntactic analysis that uses utterance simplification and intonation cues to get precise information about the syntactic structure of the Tibetan language. Native speakers, even if they are non-specialists, are thus able to provide us with precise information about the structure of utterances. This will allow the scientific community to obtain resources enabling the study of Tibetan syntax. Moreover, the extra task we have included allows for the easy generation of educational materials that the informants can benefit from.

KEYWORDS

Tibetan, NLP, syntax, word segmentation, POS tagging, Corpus Linguistics

This is a contribution from *Himalayan Linguistics, Vol. 15(1): 78–112.*
ISSN 1544-7502

Himalayan Linguistics, Vol. 15(1). © Himalayan Linguistics 2016
ISSN 1544-7502

Towards describing Tibetan syntax: From word segmentation to rewrite rules through a semi-automated workflow

Hélios Hildt
Université Bordeaux Montaigne

Like the writing systems of the Indian languages it was based upon, Tibetan is an alphasyllabic system. Tournadre describes thirty-six distinct phonemes on the basis of the phonology of modern Lhasa Tibetan (Tournadre 2014a), but the traditional presentation of 34 graphemes[1] is more relevant to our work because we will be manipulating Tibetan characters, the units of the written language. We will note them in Extended Wylie Transcription with the corresponding Tibetan characters within brackets.[2]

The writing system of a language considerably influences its capacity to be processed by NLP tools. The reason is that most of these tools are developed for the world's most-used languages, such as English, along with a few of the romance languages. Due to this simple fact, languages with writing systems that differ significantly from western languages are at a disadvantage. Among them, Chinese and Japanese do not separate lexical units, as in the case of romance languages; instead, they work at the level of the syllables.

The main obstacle to the development of tools to process Tibetan is the necessity to segment sentences into words. Since we aim at making our work usable by the average user of office suite software, we have created a segmentation tool based on a lexicon that can be updated by the user. Although an approach based on lexical resources segments correctly most of the time, it lacks the ability to deal with the segmentation of syntactic ambiguities. For this reason, it is necessary to use syntactic rules in order to determine the structure of syntagms and to look at the strings of successive Part-of-Speech (hereafter POS) tags around a sequence of syllables to determine whether it is a verb or a noun. Thus, the segmentation process depends heavily on such a syntactic analysis.

However, so far as we know, there isn't any linguistic description of Tibetan syntax that would have enabled us to improve our segmentation tool and use it as a foundation for developing a full-fledged spell-checker. It soon became obvious that without conducting proper research on Tibetan syntax, this problem will never be solved. For that reason, we have designed a semi-automated workflow that enables native Tibetan speakers to give some information regarding the structure of sentences, during our stay at Esukhia,[3] in India. We hope that the prospect of researching on syntax

1 The thirty consonantal graphemes and the four vocalic graphemes

2 This double representation ensures the readability of our work for those readers not familiar with Tibetan while reflecting our use of unicode formatted Tibetan text for all the NLP tasks we are presenting.

3 www.esukhia.org. My brother Vincent and I have created that NGO to preserve the language and literary heritage of Tibet. We offer online Tibetan lessons and lead projects that use technological advances linked to language, like NLP.

with a corpus that readily contains the structure of its utterances[4] will encourage and foster research efforts!

Furthermore, the process is designed in such a way so it only requires an extra task to provide the editor with a tangible and useful output, encouraging the informants to realise to the best of their capacities the task we present them.[5] After having provided us with the simplified structure of each and every processed sentence, the informant is just a small effort away from the simplified versions that can be directly used as educational material for students, be they first or second language learners.

1 State of the Art

1.1 Tibetan NLP

1.1.1 The "Tibetan in Digital Communication" project

With "Tibetan in Digital Communication" (hereafter TDC), Nathan Hill and Edward Garett have created the largest corpus of classical Tibetan available to the international community that is POS tagged. Thus, it is the first successful attempt at creating a corpus of such a scale where the basic NLP tasks have been carried out. The project is temporarily available at the following url: http://larkpie.net/tibetancorpus. [6] A thorough documentation of every processing involved is available, enabling one to access this monumental corpus, but also to consult the description of the tasks and find the reasons of the choices that have been made.

Ensuring the tools developed for this project are available is what really matters, from our perspective. The corpus itself will be a reference and a resource for the scientific community in general. Yet researchers specialised in Tibetan NLP will appreciate even more the ability to process their own data using the standards established by TDC within their own projects.

1.1.2 Paul Hackett

To our knowledge, besides the SOAS team that developed the TDC, the only other person in the Western academic landscape active in Tibetan NLP is Paul G. Hackett, from Columbia University, in the United States.

Using NLP techniques, he made use of a large corpus of literary works in classical Tibetan for the preparation of his Tibetan verb dictionary, which contains more than 1,700 entries. Following the trend in translation from Tibetan to Western languages that consists in consulting the Sanskrit equivalent (rather than further native Tibetan resources), he proposes the Sanskrit equivalent of each entry alongside long stretches of concordances from his corpus. The grammatical analysis he bases himself upon for the development of his verb dictionary is very close to Craig Preston's analysis, which we describe in section 1.4.3.

To our knowledge, Paul Hackett is also the only one who proposes a syntactic tree of a Tibetan sentence (Hackett 2000: 8) in a scientific paper. It is admittedly computer-generated and thus approximative, but it is a syntactic tree nonetheless.

4 The structures are derived from the simplification process that is done by native speakers.

5 A mistake in the tasks we require will result in ill-formed sentences, which will encourage them to do their very best from the beginning and to modify their work when necessary.

6 (website consulted on the 01/09/2015)

1.2 Segmentation

Various researchers have proposed solutions to the problem of segmenting sentences into words for Chinese. Many different strategies have been employed to obtain the best possible segmentation. The most complicated part is resolving the ambiguities, which are processed either by models built on lexicons and rules, or by statistical models (Lee and Huang 2013). Forward Maximal Matching (FMM) consists of looping through a string and systematically trying to match the current syllable (along with its immediate cohorts) to the longest possible word found in the lexicon; this method is extremely lightweight, which offsets the limitations inherent to any resource-based approach. Thus, if having the output of the FMM manually corrected is possible, this solution is by far the simplest to implement.

For these reasons, we have adopted this method for our segmentation tool (see section 2). Paul Hackett made the same choice for his attempt of building an Information Retrieval System for Tibetan (Hackett 2000). He describes the algorithm he used as a "greedy segmentation algorithm" that proceeds by pairing syllable sequences to a lexicon, keeping the longest possible sequence[7].

The statistical models, although much heavier in their implementation, have a greater rate of success since they can be trained on massive amounts of data. The TDC project has addressed the issue of segmentation from a different angle by considering it as the annotation of the syllables constituting a sentence within a statistical model. They have been tagged according to their position in the word: single component, first syllable, intermediary syllable or final syllable. Also, the model using Conditional Random Field (CRF) to annotate syllables had proven its worth for Tibetan, as in Liu et al. (2011). Note as well that Nathan Hill has compiled all the available works, dictionaries and lexicons about Tibetan verbs in his dictionary of verbs (Hill 2010), which is then integrated in his processing system. This has brought to light a great number of verb occurrences that would have gone unnoticed otherwise.

We do not think of these two approaches as competing against one another, but rather as different approaches applying to different contexts and meeting different needs. Although the statistical one is unarguably the most efficient and the most accurate (see section 2.4) and applies to contexts where a resource intensive system is not a problem (such as academic projects), the model based on rules and a lexicon is more suitable for being adapted for uses by the general public. In this context, accuracy is gladly traded for ease of use.

1.3 POS tagging

POS tagging enables us to fully benefit from the information our informants offer while processing utterances. Elaborating a POS tagset or evaluating those proposed by the different authors exceeds the scope of our work. For this reason, we have chosen to use the tagset of the TDC project. However, we noticed two tendencies in this domain.

The first one is represented by Tibetan authors who strive to produce a description of their language closer to the reality than what has been presented by the grammatical tradition. They start from English POS categories and amend them when it is obvious they are not relevant for Tibetan. Lobsang Monlam[8] does it for the new version of his Tibetan-Tibetan digital dictionary. Thupten Jinpa proceeds in the same way in his book on Tibetan grammar (Jinpa 2010) in which he deals with issues ignored by the tradition. As expected, his work has been strongly criticised by Tibetans, who

7 "greedy segmentation algorithm — longest substring matching to a dictionary"
8 See http://www.monlamit.org/node/20, website consulted on the 01/09/2015.

do not understand the need to innovate in a domain where the traditional approach was deemed sufficient until now.

The second tendency is represented by linguists like Nathan Hill and Edward Garett, whose tagsets seem to be the result of a linguistic analysis—or which are at least justified by linguistic considerations. We have used their work without any modification. A description of this tagset together with some examples is found at the following URL: http://larkpie.net/tibetancorpus/tags.[9]

1.4 Syntax

Tibetan syntax shines by its absence in the specialised literature: it still remains an understudied domain.[10] No study of modern Tibetan syntax is to be found; neither is there any study dealing directly with the syntax of classical Tibetan, although some correlated subjects have been investigated, like ergativity (Tournadre 1996) and anaphora (Andersen 1987).

1.4.1 Tournadre

Nicolas Tournadre talks briefly about syntax at the beginning of his career. In a paper published in 1988, he says:

> La typologie du tibétain n'est pas évidente: si la langue est flexionnelle pour un nombre limité de particules qui apparaissent fréquemment (…) elle offre des constructions agglutinantes, mais il n'en demeure pas moins qu'elle est essentiellement analytique. (Tibetan typology is no easy subject: although that language is flexional for a limited amount of recurrent particles (…) it presents some agglutinated structures while remaining nonetheless essentially analytic) (Tournadre1988: 9).

Within the description of an analytic language, it is difficult to leave aside syntax, yet, as surprising as it may seem, this domain soon disappears altogether in his research, except for some short passages about morphosyntax. As a matter of fact, he proceeds from the standpoint of morphology (Tournadre 1988) while describing the basic Tibetan syntactic models.

The typology presented by him in page 13 consists of seven "modèles structurés (du point de vue actantiel)" (models structured) (Tournadre 1988: 13).

S-V
S-O-V
S(ERG)-O(OBL)-V
O(OBL)-S-V
S-O(OBL)-V
S(ERG)-O'(OBL)-O-V

Table 1. The seven saturated syntactic models[11]

9 (website consulted on the 01/09/2015)

10 It seems to us that the absence of works on the syntax of Tibetic languages — the domain is indeed active as attested by this long bibliography — is due to the difficulty for researchers to judge the grammaticality of Tibetan utterances and to the near-absence of native speakers in the linguistic community.

11 O' = indirect object (beneficiary), = ergative, = oblique.

From this moment onwards, syntax seems to have disappeared from his research. In his *Manual of Standard Tibetan*, he only makes a few references to Tibetan syntax. (Tournadre and Dorje 2005: 395):

> The difference between Literary (Modern and Classical) and Spoken Tibetan lie in the lexicon (vocabulary), grammatical words and, to a lesser extent, syntax and pronunciation.

He adds:

> And finally, from a syntactic point of view, the written language is often more flexible than the oral. For example, adjectives and relative constructions may be placed either before or after the noun, whereas in oral Tibetan they almost always follow and precede them respectively.

Tournadre again addressed syntax in an oral communication in 2014 at the 20th Himalayan Languages Symposium, Nanyang Technological University. There, he questions the validity of the notion of subject and direct object in Tibetan. However, there is no further trace of syntax in his subsequent works.

He further says that the classical and modern varieties of Tibetan are very close to each other and that "[a]nyone who knows colloquial Tibetan can quite easily learn the literary language, and vice versa" (Tournadre and Dorje 2005: 396). The author refers first to Tibetans who learn literary Tibetan in the context of traditional religious studies and secondly to the very rare Western translators and Tibetologists who show some interest towards colloquial Tibetan.

1.4.2 Di Jiang et al.

Jiang et al. (2005) present the project titled "Grammatical Information-Dictionary of Contemporary Tibetan". In the verbs' section, the twelve groups identified at the beginning of the project are presented together with their typical syntactic structure noting the valency of each group.

1. Verbs of possession, like *yod* (ཡོད་), "possess": NP+[POS]+NP+V[12]
2. Existential verbs, like *yod* (ཡོད་), "exist": NP+NP+[LOC]+VP
3. Verbs of change, like *sgyur* (སྒྱུར་), "change": NP+VP/ +[COP]+VP
4. Perception verbs, like *mthong* (མཐོང་), "see": NP+[AG]+NP+VP
5. Directional verbs, like *'gro* (འགྲོ་), "go": NP+V, NP+VP+VP, NP+VP+[TAP]+VP
6. Cognition verbs, like *brtsi* 'jog byed (བརྩི་འཇོག་བྱེད་), "respect": NP+[AG]+NP+[OBJ]+VP
7. Narrate verbs, like *bshad* (བཤད་), "say": NP+[AG]+[NP+VP]+VP, [NP+VP]+NP+[AG]VP
8. Interrelation verbs, like *'dre* (འདྲེ་), "mix up": NP+NP+[ITP]+VP
9. Causative verbs, like *bcug* (བཅུག་), "cause to do":
   NP+[AG]+NP+[POS]+(NP+)VP+[CAU]+VP
10. Stative verbs, like *shi* (ཤི་), "die": NP+(NP+)VP

12 POS: possessive case; LOC: locative case; COP: complement particle; AG: agentive case; TAP: (no description provided); OBJ: objective case; ITP: interrelation case; DAT: dative case.

The other abbreviations are not given.

11. Action verbs are a group of verbs whose structure is: NP+VP(act), NP+NP+VP, NP+NP+DAT]+NP+VP et NP+[AG]+NP+NP+[FAT]+VP
12. Copula, like *yin* (ཡིན་), "be": NP+NP+VP

This approach and that of Tournadre both stem from empirical observations. If Tournadre is conscious of the overgeneralisations implied in the use of structures such as SOV (he eventually questions the relevance of the notion of subject (Tournadre 2014)), the Chinese researchers seem to aim at exhaustiveness and at a greater accuracy in their analysis.

1.4.3 Buddhist studies

The only ones who have dared to present a syntactic description of Tibetan are a few professors of Buddhist Studies in the United States—Magee et al. (1993), who is followed by Wilson (1992) and Preston (2005). However, they admit that their approach is essentially didactic:

> It balances traditional Tibetan grammatical and syntactic analysis with a use of terminology that reflects English preconceptions about sentence structure. Based on the system developed by Jeffrey Hopkins at the University of Virginia,(…)[13]

For this reason, it is difficult to consider it a linguistic study of syntax, eventhough they do use the syntactic representation in boxes. As we shall see it, their analysis of sentences isn't sufficiently rigorous.

The representation model of the Tibetan syntactic structure using such box-shaped diagrams was created by Jeffrey Hopkins, developed by Elizabeth Napper and Joe Wilson before being amended by Craig Preston (2005: xiv). The model presented by the latter will be analysed here. His work is a meticulous analysis of a piece of Tibetan literature and proceeds by systematically breaking down each and every sentence. It is in the introduction of this book that he describes his methodology and his didactic choices.

It is worth noting that the underlying rules of this representation are nowhere to be found. The reader can only suspect the author is basing himself on his experience of Tibetan literature to decide the way of fitting a sentence within the model we reproduce hereafter.[14]

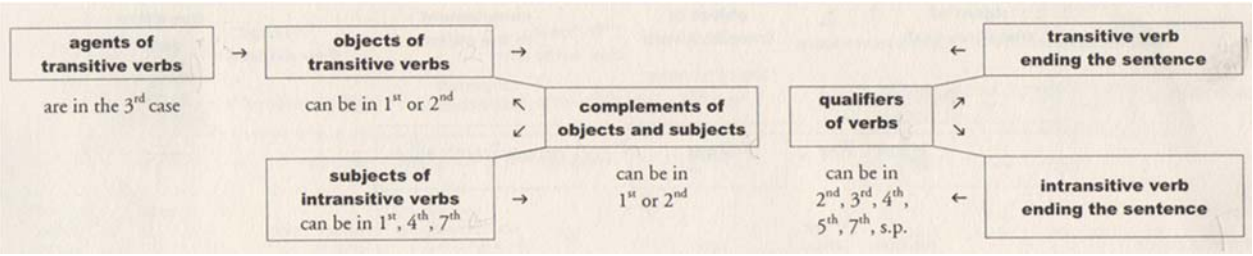


Figure 1. Preston's model of sentences

Preston's model has its limits in that it allows simple and complex sentences to fit in as long as they don't exceed the strict verbal domain. Adjuncts, thematisations and other optional

13 http://www.shambhala.com/translating-buddhism-from-tibetan.html (website consulted on the 01/09/2015)
14 The cases numbered in the image are the following: 1st case — nominative; 2th case — objective; 3th case — agentive; 4th case — benefactive; 5th case — originative; 6th case — connective; 7th case — locative; 8th case — vocative.

components are to be analysed by the student or the teacher, who will need to use his cultural knowledge and his mastery of the subject discussed. The model keeps formalisation sufficiently low in order to allow for such arrangements.

The following tree[15] was drawn using exclusively the information given by the author in regard to the analysis of the sentence that Preston translates by "In the beginning, the root of the path meets back to proper reliance on a spiritual guide; hence, you should take it to heart carefully."

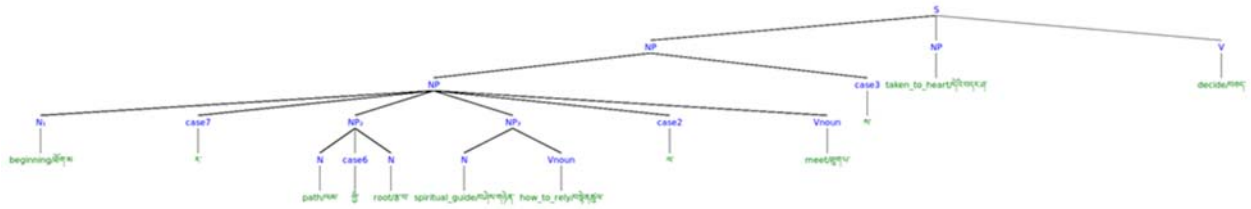


Figure 2. Syntactic tree drawn on the basis of Preston's analysis

The reader can only note that there are three nominal syntagms linked to *meet/thug* (ཐུག་), a situation that the model based on the eight constituants or eight functions does not provide for. The farthest node from the verb is linked to the sentence level in the translation made by the author, but the reader finds no explanation in regard to this change in the proposed analysis, wherein this syntagm is on the same level than the other arguments of the verb. Grouping under the same node both essential arguments and adjuncts is problematic, yet the analysis proposed by the author simply ignores it.

So, without minimising by any means the didactic qualities of this book composed for students in Buddhist Studies, those for whom a scientific rigor is a requirement can only view this work as something of an inspiration, or a mere starting point, which is what we have done in our present study.

1.4.4 Tibetan seen by Tibetans

The Tibetan grammatical tradition is of no help in our endeavour because it does not speak at all about syntax nor about grammatical categories, except for nouns, verbs, inflected particles and uninflected particles. It seems that it follows Sanskrit grammar, a free word order language where syntax plays a relatively minor role. About Parts of Speech, Sctrick says: "C'est déjà pour les Indiens un début d'inventaire structural que d'inventorier le verbe, le nom, les prépositions et les particules" (Indians already constitute a basic structural inventory [of POS] by listing the verb, the noun, prepositions and particles).

2 Segmentation

Since the development of an entirely automated segmentation tool is impossible (for the reasons discussed above), syntactic disambiguation must proceed manually by utilizing user input. Users will be taking advantage of their manual corrections by implementing their changes in our segmentation tool through the update functionalities.

15 This syntactic tree and all the following have been drawn with http://mshang.ca/syntree/.

Thanks to Élie Roux[16] who implemented in Javascript the segmentation algorithm we have adapted for Tibetan, our tool can be used both in InDesign[17] and within an internet browser.

The script is coded in plain Javascript and complies to the specifications of ECMAScript 3 (the version implemented in InDesign). No library nor any extra module has been used. The script is available at the following URL: http://eroux.fr/tibetan-wordbreak-js/[18] and the source code is available at https://github.com/EsukhiaHub/tibetan-wordbreak-js.[19] Simply copying the file in the InDesign scripts' folder is needed to have it in the list of scripts.

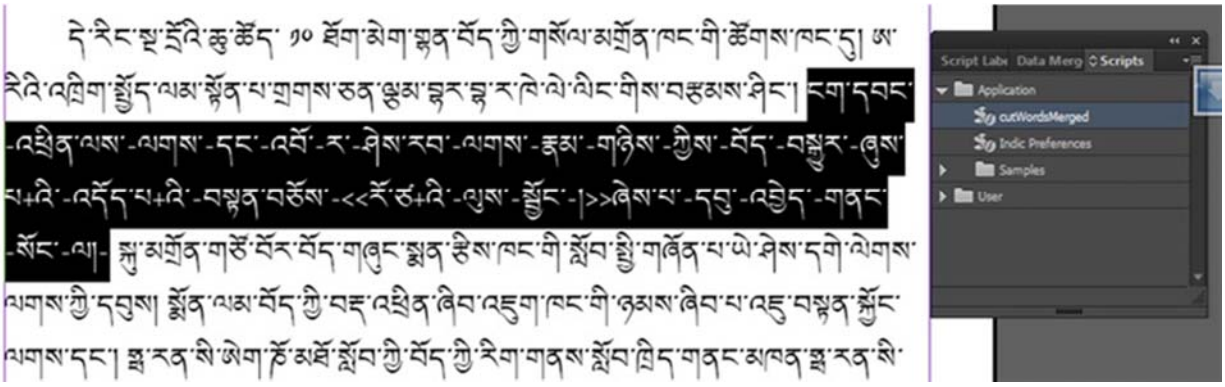


Figure 3. After running the script on a selection of text in InDesign

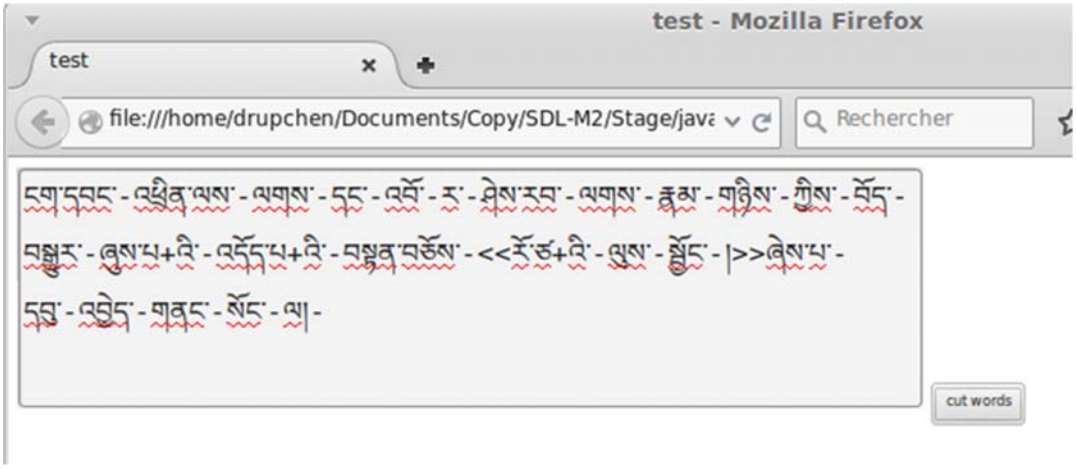


Figure 4. The same text segmented with Firefox

16 A programmer interested in Tibetan, desktop publishing software and the issues around the use of Tibetan fonts. https://github.com/eroux (website consulted on the 01/09/2015)
17 The software mainly used for publications in Tibetan language.
18 (website consulted on the 01/09/2015)
19 (website consulted on the 01/09/2015)

2.1 Algorithm

The algorithm we have implemented is Forward Maximal Matching (hereafter FMM). The size of words isn't limited in the original algorithm, but it corresponds to the length of words found in the lexical resource. Although this is the simplest approach, it soon appeared it wasn't the optimal solution for Tibetan, where the units are not characters but syllables, and where most of the words (except loan words) can be broken down, either into a combination of other words (see section 2.2) or into a combination of words and grammatical particles. Our algorithm differs from the other FMM in that it establishes a maximal length of words.

This algorithm requires:

- a list of quadri-syllabic words,
- a list of tri-syllabic words,
- a list of di-syllabic words,
- a list of mono-syllabic words,
- a list of words ending in *'a* (འ) to which *'a* (འ) was taken away,
- a list of suffixes.

2.2 Lexical resources

For the lexicon, we have chosen the Sino-Tibetan dictionary *bod rgya tshig mdzod chen mo* (བོད་རྒྱ་ཚིག་མཛོད་ཆེན་མོ་) (Yisun, 1985), the one most used in the Tibetan world[20], together with the d*ag yig gsar bsgrigs* (དག་ཡིག་གསར་བསྒྲིགས་). The latter presents the disadvantage of lacking Buddhist terms.

The first step was to isolate lemmas in each entry.

ཀ་ཀ། – ༡. སྐྱ་གའི་མིང་དང་དེའི་སྐད། ༢. ལུག་གི་ཨ་ཙུག་གམ་ཟེ་གེ། ༣. ༼ཡུལ༽ བྱིས་ཆུང་གི་གྱོན་པ།

Figure 5. Example of an entry in the Sino-Tibetan dictionary.

The second step was to organise the entries in lists according to their number of syllables. The dictionary contains compound words; however, using them for the segmentation would make us miss all the occurrences of the sequences where these rare words should be split. The easiest solution is to delete these entries from the lexical resource in order to keep only the "safe" entries (the unambiguous ones). The drawback of this solution is that we deprive ourselves of a certain number of entries. We have proceeded by successive trials, and have established that the optimal configuration is four lists of words, from mono-syllabic words to quadri-syllabic words.

20 used as a Tibetan-Tibetan dictionary, ignoring the Chinese translation

entries/lexicon	Yogācārabhūmi	Mahāvyutpatti	Rangjung Yeshe	Sino-Tibetan
1 syllable	2.34%	3.57%	2.77%	4.36%
2 syllables	12.57%	23.78%	32.91%	57.01%
3 syllables	13.11%	13.64%	19.78%	15.75%
4 syllables	23.91%	18.43%	25.09%	15.65%
5 syllables	14.78%	11.79%	7.65%	3.23%
6 syllables	12.37%	7.42%	5.52%	2.12%
7 syllables	7.94%	5.90%	2.71%	0.92%
8 syllables	4.76%	3.65%	1.46%	0.45%
9 syllables	3.09%	3.09%	0.84%	0.25%
10 syllables	1.84%	2.18%	0.42%	0.10%
10+ syllables	3.22%	6.48%	0.78%	0.11%
total entries	16030	11177	179703	52975

Table 2. The entries of four lexica ordered by length in syllables

In a previous work, we had calculated the percentages in Table 2 and noted that those entries with more than four syllables made up a negligible amount (less than 1\%).[21]

2.3 Additional functionalities

The two following fuctionalities are used to update our tool. This way, users will progressively need fewer and fewer manual corrections.

1. The personalised lists which are integrated into the lexical resources during the script's execution. They consist of three files, each containing a list of words that the user is allowed to add to:

(a) New words: the user adds the words that were not recognised by the segmentation tool, a word per line.

(b) Words to ignore: the user adds words contained in the lexical resource that create bad segmentations. The most frequent cases are entries of the Sino-Tibetan dictionary that contain inflectional particles or that are composed of concatenated words.

(c) Identified errors: the user adds errors he has identified, but that he chooses to keep[22].

2. Post-processing of the segmentation by contextual replacements. The replacements apply to the text as a whole. Each line of the file where the user registers the replacements is structured as follows:

[left context] [string to replace] [right context] [replacement string]

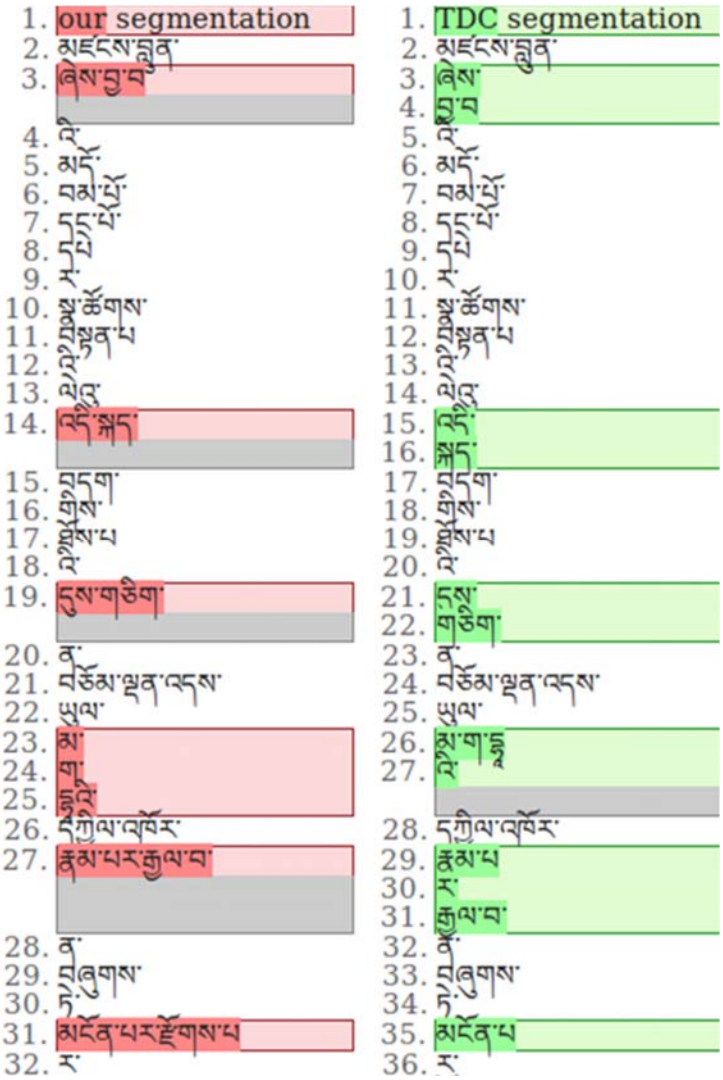
21 The lexica available at http://indica-et-buddhica.org/lexica and http://nitartha.pythonanywhere.com/search were used.

22 Used by Esukhia in the project of proofing the spelling of the Buddhist Canon in Tibetan, in the step of coming back to the spelling found in woodblocks.

The length of the strings is chosen by the user. This gives him a great freedom in establishing the degree of precision for each individual replacement.

2.4 Comparison with the TDC's segmentation

In order to evaluate the performances of our segmentation tool, we have segmented a page from the TDC project. We chose to take the raw output of their processing system.[23] We chose the raw output because, in regard to the segmentation, the only difference from the corrected version is that compound nouns are merged into single words. In most cases, these are words longer than four syllables.



23 http://larkpie.net/tibetancorpus/pretagging?textid=74&page=0, the "machine" column (website consulted on the 01/09/2015).

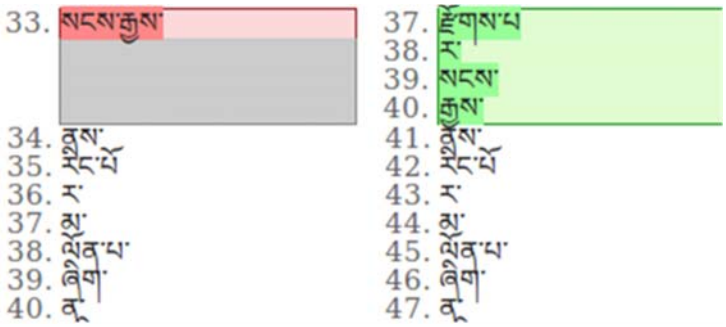


Figure 6. Comparison of our segmentation to te TDC's segmentation

Three major reasons explain the differences of the segmentation for this passage.[24] Firstly, the words in our lexical resources can still be broken into smaller units. It is the case in the first column's lines 3, 14, 19, 27 and 31. Secondly, the lines 22, 23 and 24 in the first column are a Sanskrit word not found in our lexical resource, which explains why each syllable has been considered separately and why the particle on the line 27 of the right column has not been recognised as a unit in its own right. Thirdly, the 33th line in the left column is a syntactic ambiguity. *sangs rgyas* (སངས་རྒྱས་) is a noun composed of two monosyllabic verbs, but in this context, *sangs* (སངས་) and *rgyas* (རྒྱས་) should have been separated. They should fall back to their respective original Part of Speech and constitute together the predicate of the sentence. This deconstruction is marked prosodically by a pause between these two verbs.

This comparison confirms the possibility to do word-level segmentation in Tibetan using a lexical resource-based model. It also clearly shows the next steps to improve the tool; improve the lexical resources; and continue adding contextual replacements to deal with the syntactic ambiguities individually.

3 Syntax

3.1 The two objectives of the workflow

As we have presented in the State of the Art, the domain of Tibetan syntax still remains almost completely unexplored. In order to correctly segment sentences, it is necessary to know the rules governing their formation in order to syntactically disambiguate them, however, we failed to find any study on them. We have developed a semi-automated workflow that allows us to gather information about the syntactic structure of Tibetan sentences. This process involves nothing more than native speakers using the scripts we also present in this paper.

As a matter of fact, requesting natives to explicitly state how a sentence is structured gives no result, due to the total absence of analysis about syntax in the traditional grammar. For example, this explains why Sherab (a native Tibetan speaker at Esukhia) admired our capacity to make the sentences he composed clearer simply by changing the placing of groups of words while we were helping him to translate educational material from English to Tibetan. His reaction shows his inability to consider a sentence to be made of hierarchically organised units. Not relying on mistaken intuitions about structure also ensures the result we obtain are not biased at all by our informants'

24 The entire comparison is accessible at https://www.diffchecker.com/ramfvusi (website consulted on the 01/09/2015)

representation of their language or coloured by their knowledge of some foreign language. Thus, by eliciting information on the structure of sentences rather than requesting an explanation, we are much more certain of the reliability and the quality of the data we gather.

Our workflow's output is, on the one side, the sentences cut up in units that are used to research how the sentences are structured and on the other hand, the list of progressively simplified versions of the processed sentences. This way, our informants obtain the list of all the variants of the sentences their cutting up into units has allowed the software to generate.

Simplifying the sentences in this way allows us to keep the same structure and the same words as the original, enabling the user to use the final result for educational purposes. Specifically, it is possible to generate more complex or simpler versions of the same text and use them for students of different levels.[25] Using this material, complex structures could be introduced step-by-step in order to facilitate the assimilation of the concepts presented.

Another application that will be of interest to the Tibetan world is the creation of children's stories and, eventually, the possibility to create a literature that would be adapted to each age group. There are two obstacles preventing such a literature's existence in the present day: 1) the ability to compose simple stories directly in Tibetan and 2) the ability to adapt the structure and vocabulary to various learning levels. To us, the first obstacle is the difficulty Tibetans have in producing simple utterances in writing, as literary Tibetan is traditionally reserved for formal literary domains, such as religious texts. In contrast, writing children's literature requires reflecting on the structure and vocabulary of a child's language level—in short, it actually requires a high level of skill to write low-level material. The tasks we require from our native informants in order to simplify the sentences do not require the production of new ones. Thus they are free from the need of obtaining or training for this specialisation.

Since it is a difficult, high-level skill to think of a sentence in terms of a whole constituted by smaller units and organised in a hierarchy, we circumscribe the issue altogether by simply proposing a list of all the possible sentences within which every variant of each group of words is presented. After that, the native speaker is free to either choose a sentence from the list or construct another one by picking groups of words from different sentences, using their own intuition as to which simplified version sounds best to their ear. The first obstacle (that of writing directly in an informal, low-level register) must be dealt with by a study on the vocabulary and the grammatical structures children from different ages have acquired; this falls outside the scope of the present work.

3.2 The workflow

3.2.1 Pre-processing

The first step of the workflow described above consists in segmenting the text to be processed into words. The use of our segmentation tool enabled this pre-processing to be nearly automatic. The only task we still require of our informants is to modify the segmentation when it does not correspond to the meaning of the sentence (that is, when it splits words improperly). This way, they need not

25 If the simplified story is meant to be used by beginners, they will replace the honorific forms by standard ones, the literal particles by the colloquial equivalents, and so on. The comparison of the original text with the simplified version speaks by itself: its size is significantly shrunk because it really is a simplified version and not the same story rewritten or paraphrased.

worry about particles nor about the granularity of the segmentation. They simply separate the text by inserting a new line for each sentence (pressing "enter").

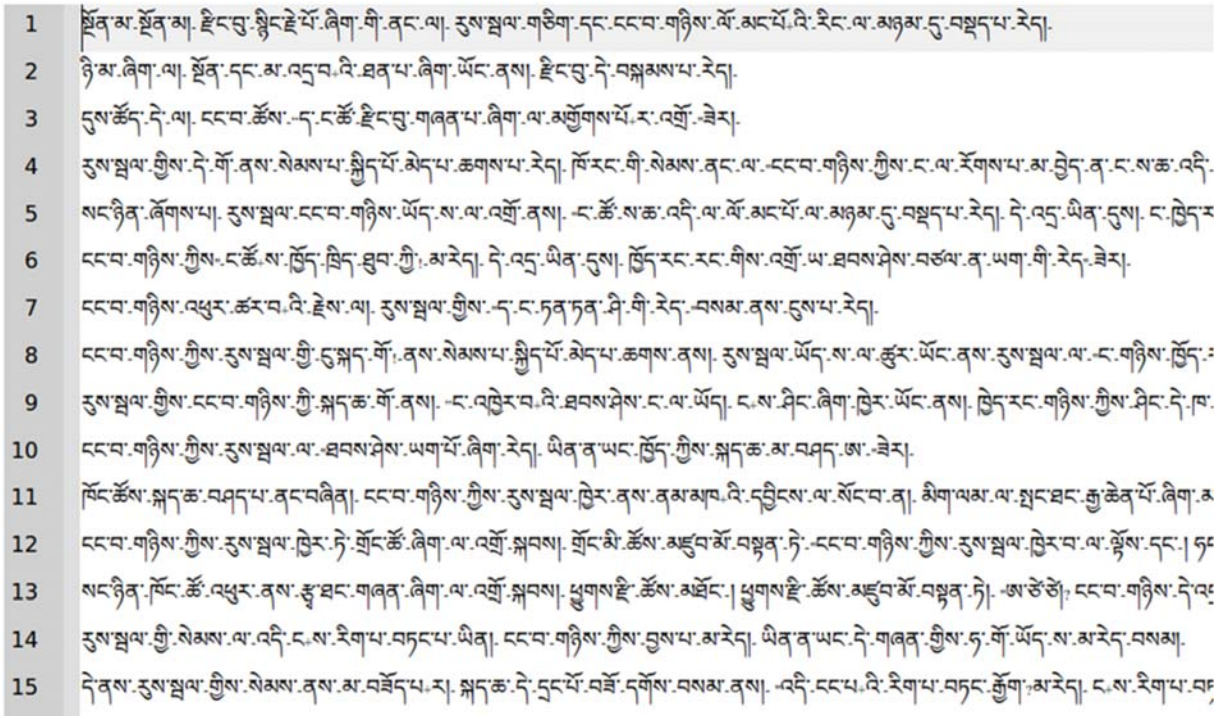


Figure 7. "La tortue et les cygnes" (The Turtoise and the Swans), segmented

Next, we create a spreadsheet containing a sentence per page and a word per cell from the segmented sentences. The name of each page is simply the number of the sentence in the text so as to have a direct access to a maximum of pages on the screen. This would not be possible if the page names would be longer, for example if they were "page 1".

The segmented sentence is inserted on the eleventh line so there is enough space to insert the corresponding image when it is available. The image permits for a greater simplification because it contains a lot of the story's contextual information. In this case, we are using the collection of stories Esukhia has digitised for which there are scanned images. This extra-linguistic contextualisation enables to reach the closest possible of the minimal structure of the sentence.

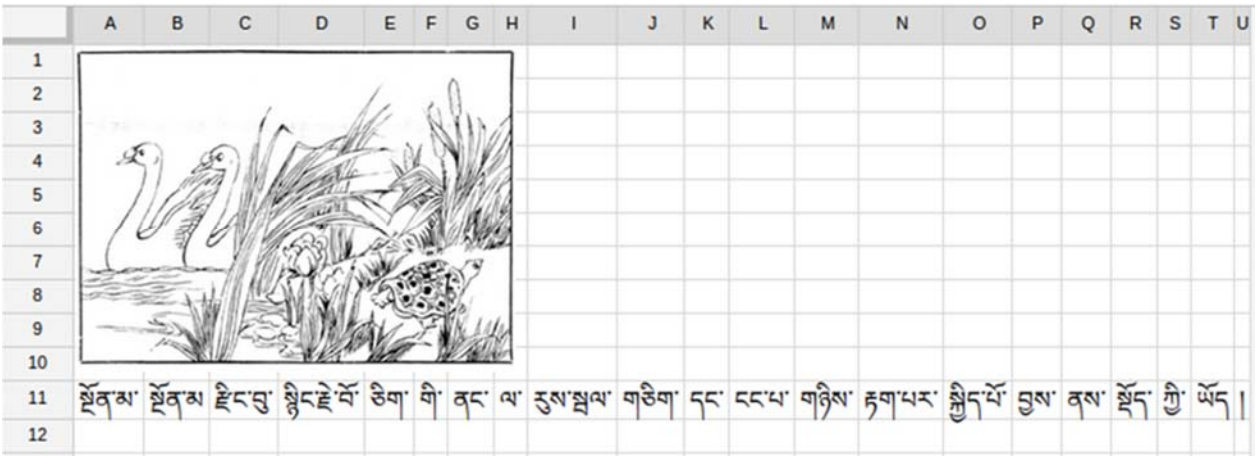


Figure 8. A sentence ready to be processed

3.2.2 Simplification

The next three steps are the most critical phases of the workflow. They bring forth the information about the syntactic structure and the validity of the generated sentences depend on them.

Here are the instructions given to our informants for this step:

1. Copy the line containing the sentence and paste it in the next line.
2. Read the sentence; look at the photo (if there is one) and take out one element in the sentence. The element may make the sentence less precise, but its general meaning must be preserved. The resulting sentence must be well-formed[26].
3. Copy and paste the new sentence into the next line; simplify again using steps 1 and
4. When nothing else can be taken out without making the sentence incorrect or incomprehensible, stop and move to the next sentence. We leave it to the native speakers to judge the grammaticality of the sentences that are each time simpler. This way, we ensure that the data we obtain will be constituted solely of grammatical sentences. This will enable non-native speakers to study the sentences' structure by relying on the proposed variants while being sure to work only on sentences whose grammaticality is established by native speakers.

26 meaning grammatical.

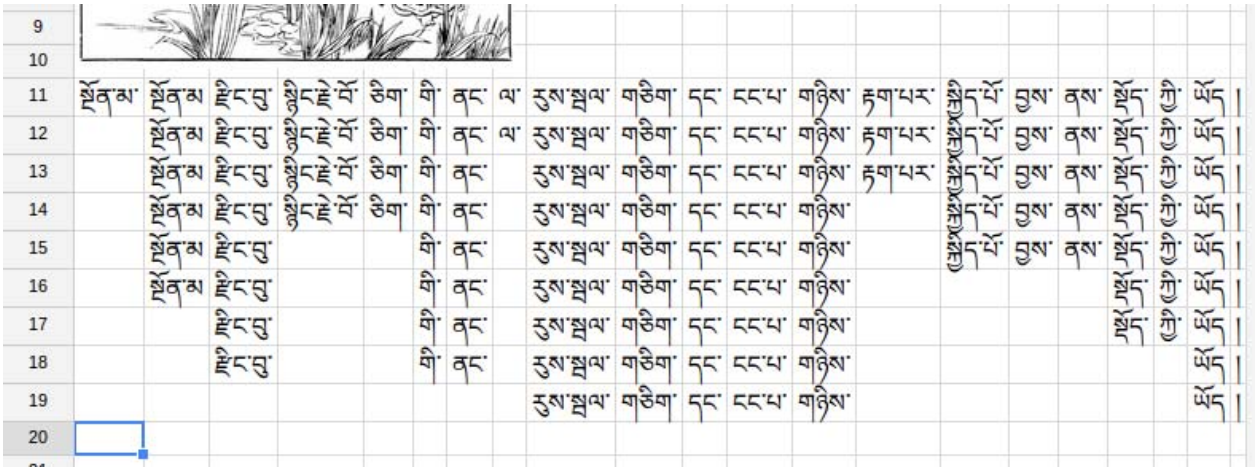


Figure 9. Simplified sentence

To illustrate what we are presenting and to clearly show the tasks we require from our informants, we will do the entire process to the following French sentence: *Ce matin, les petits enfants ont lu Babar et sont rentrés à la maison* (This morning, the young children read Babar and came back home).

Ce	matin	,	les	petits	enfants	ont	lu	Babar	et	sont	rentrés	à	la	maison	.
			les	petits	enfants	ont	lu	Babar	et	sont	rentrés	à	la	maison	.
			les		enfants	ont	lu	Babar	et	sont	rentrés	à	la	maison	.
			les		enfants	ont	lu		et	sont	rentrés				.

Table 3. Simplified sentence

We could have taken away *et sont rentrés* (and came back), but this would have modified the meaning of the sentence. *À la maison* (at home) can be taken away because *rentrer* has the meaning of *rentrer chez soi* (coming home) in case it is not followed by a complement. We stop at this fourth sentence.

The reason we stop taking out words once it doesn't make sense in the story anymore lies in our wish to only generate sentences that can enter in the composition of the simplified story without the need for rewriting. Minor adjustments are tolerated. Besides the fact that the modified sentence indicates by its own that *rentrer* means, in this particular context, *rentrer à la maison* (to come back home), we are obliged to limit to the bare minimum structural changes in the sentences for the final result to be a simplified version of the same story (and not a different story).

3.2.3 Chunking into prosodic units or sentence parts

Simplifying the sentences has given us vertical information, bringing indications about the depth of the sentences' structure. However, it is indispensable to seperate the units of meaning regrouping one or more words.

We could say this step aimed at establishing the borders of syntagms, but since enough data has not been processed and we lacked the time for a proper analysis, we prefer to use a more generic term, something like a "sentence part" or "a unit of meaning". We have first tried to explain to well-educated informants where to insert a boundary between two units of meaning, but in the best-case

scenario our request was not clear at all and in the worst-case scenario they simply didn't understand what we were asking them to do.

This observation shows that even the native speakers who have benefited from the best education are absolutely unaware of the syntactic mechanisms operating in their own language, even the most evident ones. Note that the absence of punctuation besides the optional spaces that can be inserted at the end of an idea must not help in detecting the clusters we are identifying in this step. By comparison, the boundaries that commas, colons, semi-colons, quotation marks and dashes materialise are not marked at all in written Tibetan. Their recognition is left to the reader, just as the recognition of words in the sequences of syllables that constitute sentences is left to him.

Later, we have observed that these boundaries between the units of meaning closely corresponded to the places in the sentence where it is possible to make a pause when reading it aloud. The instruction given for this step is to read the sentence aloud very slowly and to insert a blank column in the spreadsheet (to materialise the boundary) where an oral pause is possible.

Figure 10. The boundaries between sentence parts are added

Here is the spreadsheet 3 after being processed:

Ce	matin	,	les	petits	enfants	ont	lu	Babar	et	sont	rentrés	à	la	maison	.
			les	petits	enfants	ont	lu	Babar	et	sont	rentrés	à	la	maison	.
			les		enfants	ont	lu	Babar	et	sont	rentrés				.
			les		enfants	ont	lu		et	sont	rentrés				.

Table 4. Inserting the boundaries of sentence parts

Since punctuation is the convention enabling to render in writing the prosodic information needed for the correct interpretation of a sequence of words in an utterance, these bounderies will be the first ones to be captured by this task. This is what gives us the confidence that each boundary revealed to us by our informants is indeed the boundary between two units of meaning. These clusters are similar to the groups of words obtained by the shallow syntactic analysis in NLP.

3.2.4 Tagging of the indispensable sentence parts

The last step of this set consists of tagging the parts of the sentence required to maintain its grammaticality and to ensure it remains intelligible in context.

On one hand, this tagging enables us to identify the parts of the sentence that play an important role in the sentence, such as the predicate and its essential arguments. The adjuncts and the optional arguments are identifiable in contrast to these. On the other hand, tagging is required for the next step; generating simplified versions of the sentences. This is because we need to know which parts must be kept and which parts can be omitted. The instruction we give here is to highlight, in yellow, the parts of the sentence that are required to maintain a correct sentence that still contains all the information needed for the story to make sense.

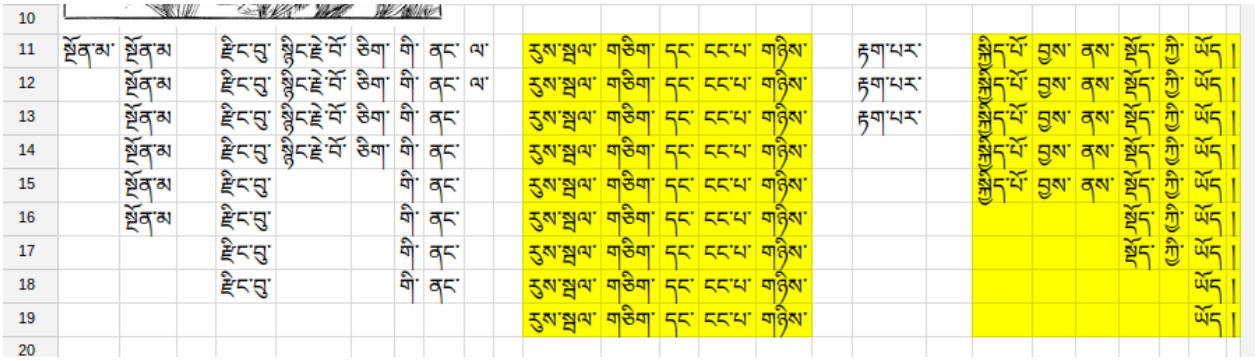


Figure 11. The indispensable parts of the sentence are highlighted in yellow.

By highlighting in yellow the indispensable parts of the table, here is what we obtain:

Ce	matin	,	les	petits	enfants	ont	lu	Babar	et	sont	rentrés	à	la	maison	.
			les	petits	enfants	ont	lu	Babar	et	sont	rentrés	à	la	maison	.
			les		enfants	ont	lu	Babar	et	sont	rentrés				.
			les		enfants	ont	lu		et	sont	rentrés				.

Table 5. The indispensable parts are marked in red.

3.2.5 Marking of the nested sentences

In order not to saturate RAM memory while generating the sentences, we request our informants to mark the boundaries between two nested sentences constituting a complex sentence. When the number of words in a sentence exceeds approximately twenty, they are asked to mark in blue the boundary between the nested sentences, usually linked by a grammatical unit.

Figure 12. The complex sentence is split.

3.2.6 Post-processing: Generating the sentences and the table of possible parts

During the steps of simplification (3.2.2), establishing the boundaries between parts of sentence (3.2.3) and the marking of nested sentences (3.2.4), we obtain the data which compose our resource for further research on syntax. The next steps only aim at producing the educational material justifying that native speakers go through the hassle of processing our data.

In this step, we generate all possible correct sentences allowed by the information gathered in the previous steps. The input for the tool that generates the sentences is a copy-paste of the table together with a list of the numbers corresponding to the yellow parts of the sentence. For example, for the sentence in Figure 11, this list will be the following: "[3, 5]". The person executing the script will copy in the sentence generation tool either the entire sentence, or the nested sentences one at a time. Doing so allows for a significant reduction of the number of generated sentences. Far too many sentences — about three hundred — would be generated for a sentence exceeding thirty words if it were to be processed in a single block.

The tool gives three elements as its output:

1. The list of possibilities for each sentence. (The words within each part of sentence have been concatenated while the sentence parts remain separated).
2. The number of generated sentences.
3. The generated sentences, one per line.

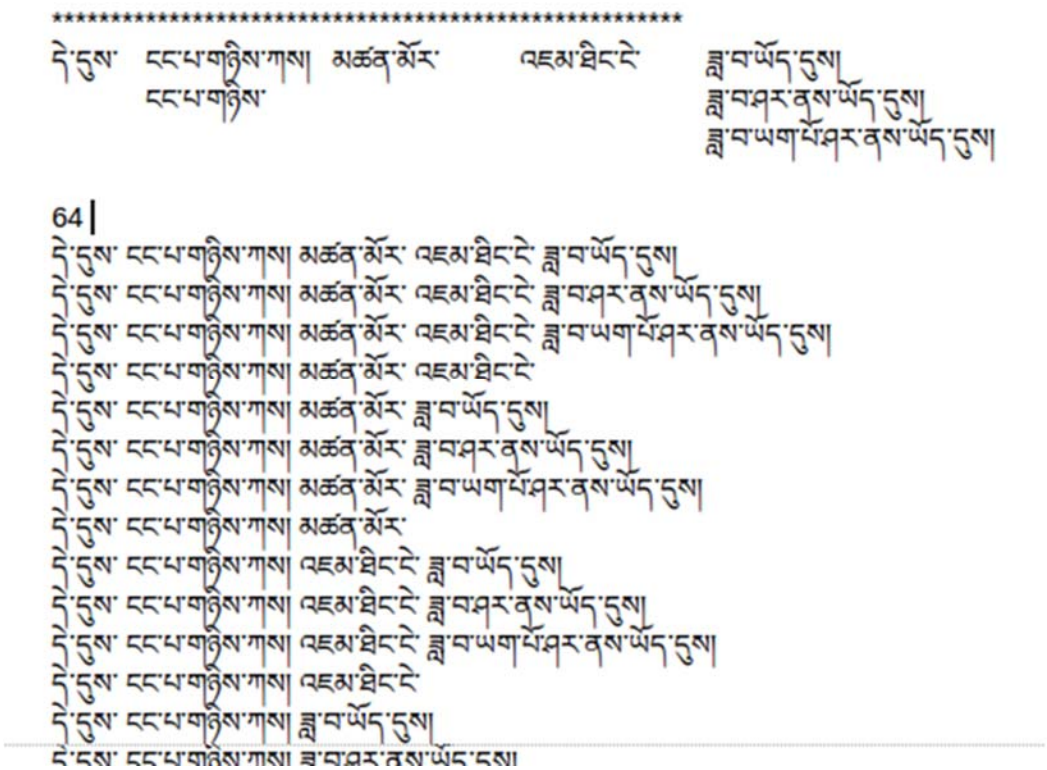


Figure 13. The first sentences generated for Fig. 11

Here are the sentences produced by the sentence generation tool for our French example:

16
ce matin, les petits enfants ont lu Babar et sont rentrés à la maison.
ce matin, les petits enfants ont lu Babar et sont rentrés
ce matin, les petits enfants ont lu et sont rentrés à la maison.
ce matin, les petits enfants ont lu et sont rentrés
ce matin, les enfants ont lu Babar et sont rentrés à la maison.
ce matin, les enfants ont lu Babar et sont rentrés
ce matin, les enfants ont lu et sont rentrés à la maison.
ce matin, les enfants ont lu et sont rentrés
les petits enfants ont lu Babar et sont rentrés à la maison.
les petits enfants ont lu Babar et sont rentrés
les petits enfants ont lu et sont rentrés à la maison.
les petits enfants ont lu et sont rentrés

les enfants	ont lu	Babar	et sont rentrés	à la maison.
les enfants	ont lu	Babar	et sont rentrés	
les enfants	ont lu	et sont rentrés	à la maison.	
les enfants	ont lu	et sont rentrés		

Ce matin,	les petits enfants les enfants	ont lu	Babar	et sont rentrés	à la maison.

Table 6. The generated sentences

1 བློ་མ་ཞིག་གིས་རྒྱུན་དུ་ཞོགས་ཤར་ཞལ་དང་སྤྱན་བསྡུམས་ཏེ། ཅོག་ཅོག་གནང་ནས་ཆུ་ཚོད་ཕྱེད་ཀ་ཤྱག་ ཙམ་བཞུགས་ཀྱི་ཡོད།
1 བློ་མ་ཞིག་ཁ་དང་མིག་བསྡུམས་ནས། ཅོག་ཅོག་བྱས་ནས་བསྡོད་ཀྱི་ཡོད།
2 ཉིན་གཅིག་ཕྱུག་ཞིག་གིས་བློ་མ་ཁོང་དེ་ཤྱར་གནང་བ་མཐོང་རྗེས།
2 ཕྱུག་ཞིག་གིས་བློ་མ་དེས་དེ་འདྲ་བྱས་པ་མཐོང་ནས།
3 བློ་མ་ལགས། ཕྱིད་རང་དེ་འདྲ་བཞུགས་ནས་ག་རེ་གནང་གི་ཡོད། ཅེས་དྲིས་པར་
3 "ཕྱིད་རང་དེ་འདྲ་བསྡོད་ནས་ག་རེ་བྱེད་ཀྱི་ཡོད།"
4 བློ་མ་དེས། ང་ཁ་འདོན་བྱེད་ཀྱི་ཡོད། ཅེས་གསུངས་པ་རེད།
4 བློ་མ་དེས། "ངས་ཁ་འདོན་བྱེད་ཀྱི་ཡོད།" ལབ་པ་རེད།
5 ཡང་ཕྱུག་དེས། ཨ་ལས་ཞལ་བསྡུམས་ནས་བཞུགས་ན་ག་འདྲས་ཞལ་འདོན་གནང་གི་ཡོད།
5 ཕྱུག་དེས། "ཁ་བསྡུམས་ནས་ཁ་འདོན་ག་འདྲས་བྱེད་གི་ཡོད།"
6 ཞེས་དྲིས་པར་བློ་མ་དེ་ཅི་ཟེར་འདི་ཟེར་མི་ཤེས་པར་ལྷས་སོ། །
6 བློ་མ་དེས་ག་རེ་ལབ་ཏོས་ཏ་མོ་བ་ཚགས་པ་རེད།

Figure 14. Comparison of the original story with its simplified version

3.2.7 Assessment of the workflow

The instructions we have proposed were put together in collaboration with the informants working at Esukhia, which ensures that they are followed correctly in their current formulation. The one for segmenting in words, "go through the segmented text and change the misplaced segmentations", does not raise any problems, and the segmentation obtained by our tool does not leave ambiguous cases. The only segmentations requiring an action are those which make the sentence ungrammatical. They are easily detected and corrected. The instruction to copy the sentence, too, does not raise any issue.

The other instructions, concerning the deletion of words, usually raises some questions. It ought to be carefully explained that there is a core message of a sentence that is modified by additional words, and that these additional words may be removed if their deletion does not change this core meaning, or make the story not understandable. Once these points are clear, the instructions are easily followed perfectly.

Since the results we obtained have lived up to our expectations, we have not felt the need to reformulate the instructions, or add more details (for example by requiring a particular order to be

followed for word deletion). Though each informant proceeds in his own way, the results from two different informants differ very little in the end. If there is a weak point in the workflow's tasks, it is in the instructions to add boundaries between sentence parts. "Read very slowly" can easily be interpreted in such a way that the number of added boundaries is insufficient. For now, this shortcoming is addressed easily enough by training shortly with the informant, and processing a few sentences together.

Even so, we often end up with too few separations; on the other hand, these separations have always been correctly positioned. We have never seen a boundary positioned in an implausible location, like inside a nominal syntagm.

The instruction about highlighting in yellow the indispensable parts of a sentence did not raise any problem. We observed a correlation between the last parts of sentences and the highlighted parts, when there were no nested sentences. However, within complex sentences, the informant considers the sentence as a whole and does not highlight the parts within the optional nested sentences. The person running the sentence generation tool should either be able to find the indispensable parts within the nested sentences or have a native speaker help him do so.

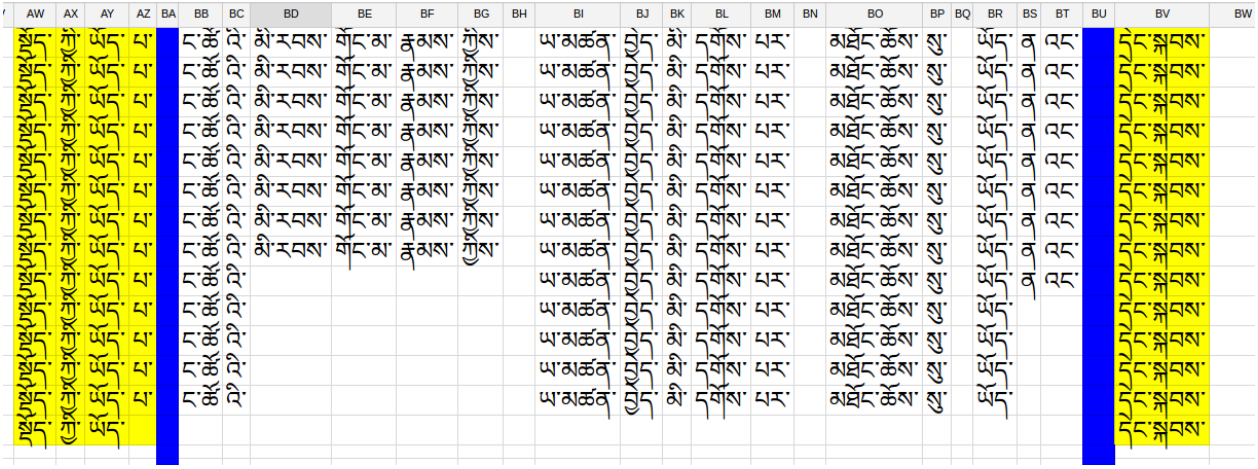


Figure 15. Example of a nested sentence having received no highlight

The direct relationship between the quality of the data we gather and the result our informants get from the workflow ensure we always obtain reliable data.

4 The syntactic analysis

The content of this section should be regarded as one of the many possible outcomes from using the structured sentences obtained from the workflow we present. We have chosen to stick as closely as possible to the tree structures our mini-corpus of three processed sentences brought to light in order to show the extent to which the output of our workflow is useful.

Since creating dependency structures implies determining the agents for each sentence while our workflow does not provide this information, we have chosen to build constituency syntactic trees.

4.1 Data preparation

4.1.1 POS tagging

The POS tags from the TDC are almost systematically constituted of two parts separated by a dot and are exclusively in lowercase. The left part is the general category and the right part is a subset of the first category. Dividing the tag with a dot allows for very elaborate tags while keeping its readability and allows to dispense with unclear terminology. This is how *n.v* indicates that the designated noun is built from a verb, *n.v.past* indicates that the past form of the verb has been used in the nominalisation process.

POS	Definition
adj	adjectives
adv.temp	temporal adverbs
case.abl	ablative case
cl.focus	Focus clitics
cv.abl	ablative converb
d.dem	demonstratives
interj	interjections
n.count	count nouns
n.v.aux	Nominalization of auxilliary verbs
n.v.cop	Nominalization of copulas
neg	negation
num.card	cardinal numbers
p.pers	personal pronouns
punc	punctuation
v.cop	Copula verbs

Table 7. A selection of POS tags from the TDC project. The complete set is annexed.

We have made use of this clever strategy in our notation of syntactic trees. We have attempted to limit subjectivity as much as possible in passing from POS tagged data to the syntactic trees. To that end, we have attributed the tag *"S"* to the syntagms composed of a unit ending by a case particle[27]. Following this convention, the syntagm composed of a noun ending by the agentive case particle will be tagged *S.agn*. This working convention enables us to benefit from a coherent notation without needing to study in details the theorical choices made by the authors of the TDC. This way, we have distinctive tags for all the cases we will encounter that don't bear the marks of any theorical analysis besides the one underlying the POS tagset we already use.

27 At places where the data we gathered from our workflow only gave us big chunks of text with no further hints at how they should relate to each other, we have relied on other parts of our small corpus where similar structures were broken down. For all the cases not covered by our data, we chose the simplest possible structure with the blocks at hand. For example, two noun phrases with dang (དང་) in between has been analysed as [(noun phrase) (dang (དང་)) (noun phrase)].

Two adjustments of the POS tagset from the TDC project were necessary to exploit our data. Firstly, we have split apart the *n.v* located in the verbal syntagms that receive the Tense-Aspect-Mood affixes. Separating the verb from the nominaliser *pa* (པ་) and its contextual variants allows us to keep the boundaries proposed by our informants, who asked us expressly to keep them apart in the segmentation in words. The other occurrences of nominalisations have been left as they were because they functioned on the syntactic level as distinct units.

Secondly, we have grouped the disyllabic verbs constituted of two monosyllabic verbs because the meaning of the composed verb is more than the mere concatenation of the two monosyllabic verbs. Without grouping the two parts of the verb, we face the incapacity to produce a satisfactory gloss.

4.1.2 Gloss

On top of the POS tags, we have glossed the story in an approximate way, since the POS tags already brought a finer analysis. We propose a French translation of each unit and we only resort to the abbreviations listed below when a unit can not be directly translated. Readers familiar with Tibetan can ignore the gloss altogether, as it is only provided as a way to understand the meaning of the sentence being analysed.

GI: agentive case particle realised by one of the inflected forms of *gi* (གི་).
LA: locative or dative case particle realised by one of the inflected forms of *la* (ལ་).
NOM: a particle that nominalises the preceding unit.
EXCL: exclamative particle or exclamative expression.
NEG: negation particle.
AUX: any part of the verbal word except the verbal root.
PONCT: punctuation symbol.

4.2 From spreadsheets to syntactic trees

The next step in the exploitation of our data consists in reaching the higher level of abstraction, the syntactic tree representation. We give to the *S* tag found in nodes a slightly different meaning. It represents to us the parts of sentences that have been delimited by our informants. We chose this convention because it is the most natural for tags on the level of nodes and because it does seem that the established boundaries are syntagm boundaries.

Some remarks in regard to the drawing up of trees:

- We have grouped under a node starting with *S* the units evidently composing a syntagm, like nominal syntagms that we also note *S.n*, to maintain the consistency of our notation. In the cases where the analysis isn't readily provided to us, we limit ourselves to grouping the units under a node. For example, the last part of the second sentence is indeed a verbal syntagm, but its inner structure exceeding what can be analysed at this point, we have left all its components on the same level. A corpus larger than these three sentences will certainly allow a better analysis.
- The last case particle found at the end of a sentence part or at the end of a syntagm (case.xxx) gives its name to the node in which it is found.
- In a node ending with a case particle, we have grouped in a syntagm the units that are in its domain of governance. Most of the time, this covers everything preceding the case particle on a single level.

- A given node can give its name to the node of the higher level when it is complemented by another node. Thus, for the structure *[S.n] [S.v]*, given that *S.n* is governed by *S.v*, we group both of them under a *S.v*. The complete structure is *[S.v [S.n] [S.v]]*.

In order to generate the collection of trees, we have proceeded in two steps. First, we have progressively reconstituted the complete tree structure following a bottom-up order, starting on the last sub-tree, where the general structure is the clearer. Secondly, from the global structure, we have generated the different trees following a top-down order. It is the same order our informants have followed when they progressively suppressed the optional words.

Proceeding in two steps enables us to take into account the general architecture of the sentence while drawing the simplified trees. We have also rendered the optional parts of the sentence (the ones not highlighted) by putting the node tag into brackets. They are placed on the level directly inferior to the sentence level. The nodes that are obviously optional have also been put into brackets.

Finally, note that there are occurrences where the grammatical categories in our gloss and those in the POS tags don't correspond. The gloss being meant for the readers unfamiliar with Tibetan, it hesitates between the words found in our translation and a more literal translation.

4.3 Examples of progressively complex syntactic trees

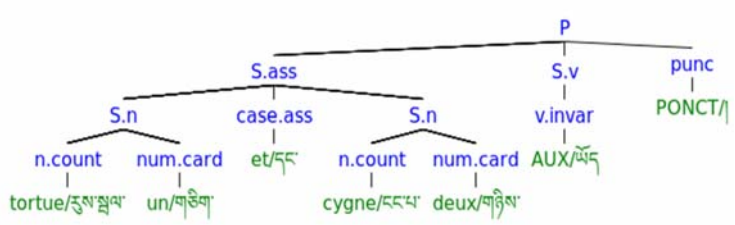


Figure 16. First tree of sentence 1

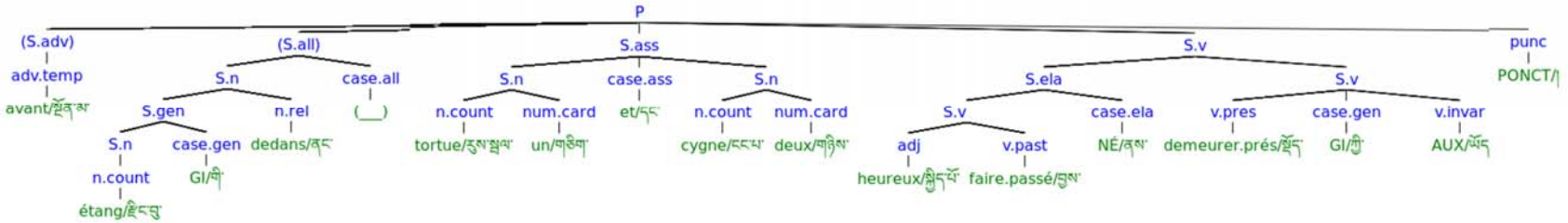


Figure 17. Fourth tree of sentence 1

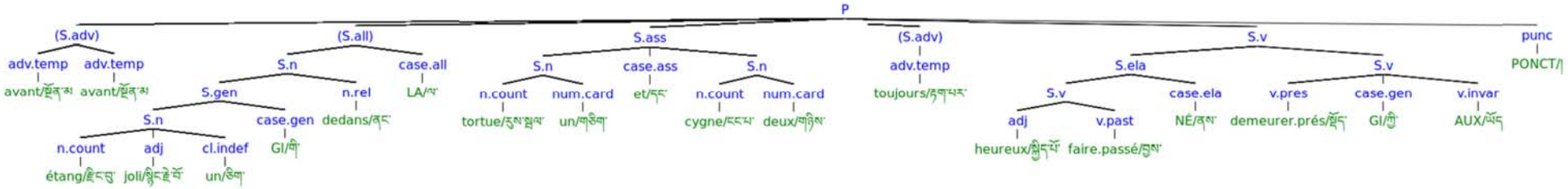


Figure 18. Ninth tree of sentence 1

Himalayan Linguistics, Vol. 15(1). © Himalayan Linguistics 2016
ISSN 1544-7502

4.4 A sketch of the rewrite rules

Once the syntactic trees have been generated, we were able to establish the list of all encountered structures and to establish their respective frequency.

In order to identify all the structures found in the first three sentences, we have browsed each tree, from the simplest to the most elaborate and we have listed the composition of each note we encountered. Later, we have counted the number of occurrences of those rules by browsing the complete tree of each of the three sentences. This is how we have been able to establish the most frequent structures, those capable of pertaining to the highest level of generality within the scope of this very limited corpus.

	OCCURENCES	REWRITE RULES
(*1*)	2	P → (S.all) S.n S.v (punc)
(*2*)	1	P → (S.adv) (S.all) S.ass (S.adv) S.v punc
(*3*)	1	P → (S.ela) S.all S.v
(*4*)	1	P → (S.n) S.agn punc (S.adv) S.ela S.v punc
(*5*)	1	P → (S.term) S.ela punc
(*6*)	1	P' → P P
(*7*)	4	S.ela → S.v case.ela
(*8*)	2	S.all → S.n (case.all)
(*9*)	2	S.gen → S.n case.gen
(*10*)	2	S.term → S.n case.term
(*11*)	1	S.agn → S.n case.agn
(*12*)	1	S.all → S.v case.all
(*13*)	1	S.ass → S.n case.ass S.n
(*14*)	1	S.ela → S.n case.ela
(*15*)	1	S.focus → S.n cl.focus
(*16*)	2	S.adv → (adv.temp) adv.temp
(*17*)	1	S.adv → adv
(*18*)	1	S.adv → P adv.temp
(*19*)	3	S.n → n.count num.card
(*20*)	2	S.n → n.count (adj)
(*21*)	1	S.n → d.dem n.count

	OCCURENCES	REWRITE RULES
(*22*)	1	S.n → n.count (S.focus)
(*23*)	1	S.n → n.count adj cl.indef
(*24*)	1	S.n → n.count adv
(*25*)	1	S.n → n.count p.indef
(*26*)	1	S.n → n.v.fut
(*27*)	1	S.n → num.card
(*28*)	1	S.n → S.gen n.count (adj)
(*29*)	1	S.n → S.gen n.count d.dem
(*30*)	1	S.n → S.gen n.rel
(*31*)	1	S.n → S.n d.dem
(*32*)	2	S.v → S.n S.v
(*33*)	1	S.v → adj S.v
(*34*)	1	S.v → adj v cl.tsam v.past
(*35*)	1	S.v → adj v.past
(*36*)	1	S.v → cl.neg v nom
(*37*)	1	S.v → P S.v
(*38*)	1	S.v → S.ela S.v
(*39*)	1	S.v → S.ela v.pres
(*40*)	1	S.v → S.n v.fut v
(*41*)	1	S.v → S.n v.past
(*42*)	1	S.v → S.v case.gen v.cop
(*43*)	1	S.v → v.pres
(*44*)	1	S.v → v.pres case.gen v.invar

Table 8. The rewrite rules

4.5 Confrontation of our results

4.5.1 Tournadre

By comparing the basic models of Tournadre reproduced in Tab. 1 to the rewrite rules of our three sentence corpus, we have found the same structure on the top of the list: *S-V* that correspond to our *(1)* rule. Then, *S-O-V* also corresponds to the *(1)* rule, and the next, *S( )-O-V*, corresponds to the *(4)* rule, provided *S.ela* is ignored.

4.5.2 Preston

Preston offers an analysis that is more interesting to compare to our results. We have chosen for that comparison the title of the literary work described in Preston (2005 : xxvii-xxxii) to ensure we did not choose a sentence that didn't receive all the necessary attention from the author.

The presentation of this gloss is ours, all the implied analysis is from Preston.

(1) སྐྱེ་བུ་ གསུམ་ གྱི་ ཉམས་སུ་ལྡང་བ་ འི་ རིམ་པ་ ཐམས་ཅད་ ཚང་བར་ སློབ་པ་
N Adj case3 Vnoun case6 N Adj Adv Vnoun
beings three 3rd practice 6th stages all completely teach
འི་ བྱང་ཆུབ་ (ཀྱི་) ལམ་ གྱི་ རིམ་པ་ བཞུགས་སོ་
case6 N (case6) N case6 N V
6th enlightenment (6th) path 6th stages live+SYNTACTIC PARTICLE
Stages of the path to enlightenment thoroughly teaching all the stages of practice by the beings of three [capacities].

The author provides us with an analysis of this title spanning over six pages. Every part of the sentence is detailed at varying levels of precision. Everything needed to build syntactic trees corresponding to his analysis is found in these pages.

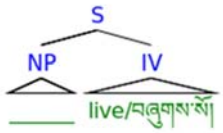


Figure 19. Sentence level

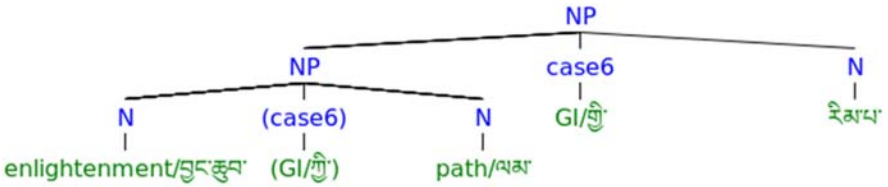


Figure 20. First NP

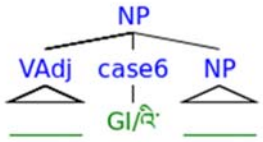


Figure 21. Second NP

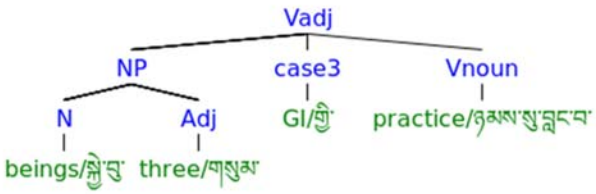


Figure 22. First VAdj

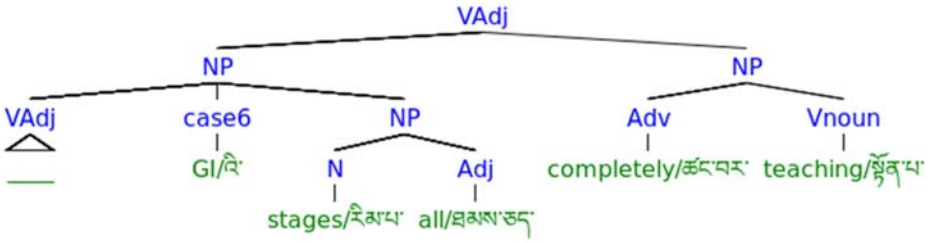


Figure 23. Second VAdj

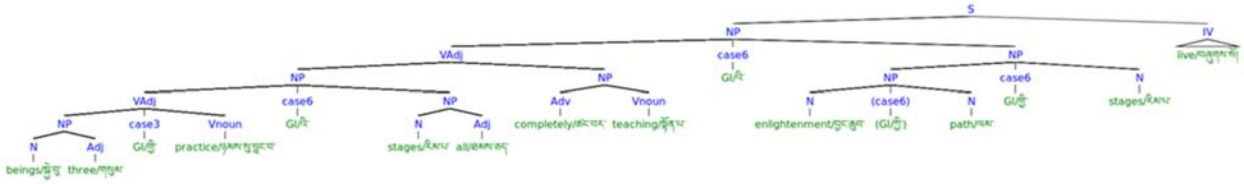


Figure 24. The entire sentence

	REWRITE RULE
(*1*)	S → NP IV
(*2*)	NP → NP case6 N
(*3*)	NP → N (case6) N
(*4*)	NP → VAdj case6 NP
(*5*)	VAdj → NP case3 Vnoun
(*6*)	VAdj → NP NP
(*7*)	NP → VAdj case6 NP
(*8*)	NP → N Adj
(*9*)	NP → Adj Vnoun

Table 9. Rewrite rules derived from Preston

The comparison of the two tables of rewrite rules shows a great similarity between ours and Preston's analysis. His rule (1) corresponds to our (1) — similar to Tournadre —, his rules (2) and (3) corresponds to our (28). The remaining rules show what Preston analyses as a verbal adjective (*VAdj*). Basing ourselves on the TDC tagset, we analyse verbs followed by a nominaliser as verbal nouns. The other rules, like the (6), seem fairly impossible, more investigation would be required to confirm them.

4.5.3 Di Jiang et al.

Concerning the sentence patterns corresponding to the different types of verbs presented by Jiang Di and his collaborators, the first corresponds once again to our first rewrite rule, but their analysis covers too large a spectrum of cases to be compared to our three sentence corpus. The only thing we can say is that the sentence structures they present seem to be compatible with our analysis.

As shown in this section, our experimental approach makes it possible to initiate an empirically driven study of Tibetan syntax, regardless of whether we are native speakers or not.

5 Conclusion

We have implemented a FMM segmentation algorithm, adapting it to the needs of Tibetan and adding to it the functionalities of user-side update. The counterpart of being lightweight and being executed in InDesign is that it does not deal with syntactic ambiguities and its quality is entirely dependent on the quality of its lexical resources. Even if it does give us complete satisfaction on being light, showing it could be integrated in office suites, the lexical resources still need to be improved. The entries of the Sino-Tibetan dictionary containing particles or constituted of composed words still need to be detected and separated. From that point on, limiting the maximal length of words won't be necessary anymore.

The semi-automated workflow we have designed enabled us to propose a fairly precise syntactic analysis of Tibetan utterances. The two simple tasks we require from our informants can be executed by any native speaker, as we have experimented with at Esukhia. First, we ask them to progressively suppress the words in an utterance that do not affect its grammaticality nor the message it conveys. Then, we ask them to read aloud the utterance very slowly and to insert blanks where they can make a pause. The parts of each utterance is indicated to us in this way.

The data we gather are presented in a spreadsheet containing the generated versions of the utterance that are progressively simplified. We then proceed to POS Tagging. We replace the words in the spreadsheet by the tags. Finally, the vertical information (the parts of the utterance) and the horizontal one (the inner structure of these parts) is transposed into syntactic trees. The comparison of all the trees enables us to establish the list of rewrite rules pertaining the analysed corpus. Taking into account all the different versions of each utterance that are processed, the results we obtain with a corpus limited to three sentences can already compare with the state of the art.

For their part, our informants get a list of all the grammatical sentences that can be generated from the information obtained through the two tasks we require. At that point, they can choose the variant whose level of complexity corresponds to their learning objectives and obtain a version of the original text containing the same sentences that are simplified.

References

Anderson, Stephen A. 1992. *A-morphous morphology*. Cambridge: Cambridge University Press [Cambridge Studies in Linguistics 62].

Andersen, Paul Kent. 1987. "Zero-anaphora and related phenomena in classical Tibetan." *Studies in Language* 11.2: 279-312.

Hackett, Paul G. 2000. "Automatic segmentation and part-of-speech tagging for Tibetan: A first step towards machine translation." In: *Proceedings of the 9th Seminar of the International Association for Tibetan Studies*. Leiden: The Netherlands. URL: http://hdl.handle.net/10022/AC:P:10471 .

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

Jiang, Di; Long, Congjun, and Zhang, Jichuan. 2005. "The verbal entries and their description in a grammatical information-dictionary of contemporary Tibetan." *The 2nd International Joint Conference on Natural Language Processing*, 874-884. Berlin; Heidelberg: Springer-Verlag [Asian Federation of Natural Language Processing].

Lee, Chia-Ming; and Huang, Chien-Kang. 2013. "Context-based Chinese word segmentation using SVM machine-learning algorithm without dictionary support". In: *Sixth International Joint Conference on Natural Language Processing*, pp. 614-622.

Liu, Huidan, et al. 2011. "Tibetan word segmentation as syllable tagging using conditional random field". In: *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pp. 168-177.

Magee, William A. et al. 1993. *Fluent Tibetan: A proficiency oriented learning system (Novice and Intermediate Levels)*. Ithaca, NY: Snow Lion Publications.

Preston, Craig. 2005. *How to read Classical Tibetan, Volume One: Summary of the general path*. Snow Lion Publications.

Sctrick, Robert. "Parties du discours" (Parts of Speech). *Encyclopedie Universalis.*
URL: http://www.universalis.fr/encyclopedie/parties-du-discours/.

Thubten Jinpa. 2010. *Bod skad kyi brda sprod gsar bsgrigs smra sgo'i lde mig* (The Key of Eloquence, a New Tibetan Grammar). Archana.

Tournadre, Nicolas. 1988. "A propos du sujet et de la morphologie nominale en tibétain: Vision traditionnelle et proposition d'outils descriptifs modernes" (About the subject and nominal morphology in Tibetan: Presentation of the tradition and a proposition of modern descriptive tools). *Bulletin de la Société de Linguistique de Paris* 83: 277-292.

Tournadre, Nicolas. 1996. *L'ergativité en tibétain: approche morphosyntaxique de la langue parlée* (Tibetan ergativity. A morphosyntactic approach of the Spoken Language). Louvain: Peeters Publishers [Bibliotheque de l'Information Grammaticale].

Tournadre, Nicolas. 2014a. "The Tibetic languages and their classification." In: Owen-Smith, Thomas; and Hill, Nathan W. (Eds.) *Trans-Himalayan linguistics, historical and descriptive linguistics of the Himalayan area*, 105-129. Berlin: Mouton de Gruyter.

Tournadre, Nicolas. 2014b. "Arguments against 'subject' and 'direct object' as viable concepts in Tibetan: revisiting the syntactic categories suitable for Classical Tibetan as well as the modern Tibetic languages". Paper presented at The 20th Himalayan languages Symposium. Nanyang Technological University, Singapore.

Tournadre, Nicolas; and Sangda Dorje. 2005. *Manual of Standard Tibetan: Language and civilization.* Ithaca: Snow Lion Publications.

Tsetan Shabdrung. *Bod gangs can gyi sgra rigs pa'i bstan bcos le tshan 'ga' phyogs bsdus (A few points compiled from the grammatical tradition of the Land of Snow).* Xining: Mtsho sngon mi rigs dpe skrun khang (Kokonor People's Publishers).

Wilson, Joe B.. 1992. *Translating Buddhism from Tibetan: An introduction to the Tibetan literary language and the translation of Buddhist texts from Tibetan.* Ithaca: Snow Lion Publications.

Zhang, Yisun. 1985. *Bod rgya tshik mdzod chen mo*(The great Sino-Tibetan dictionary). Beijing: Mi rigs dpe skrun khang (People's Publishers).

Hélios Hildt
hhdrupchen@gmail.com

APPENDIX: POS TAGS FROM TDC

POS	Definition
adj	adjectives
adv.dir	directional adverbs
adv.intense	intensive adverbs
adv.mim	mimetic adverb
adv.proclausal	proclausal adverbs
adv.temp	temporal adverbs
case.abl	ablative case
case.agn	agentive case
case.all	allative case
case.ass	associative case
case.comp	comparative case
case.ela	elative case
case.gen	genitive case
case.loc	locative case
case.nare	the case -na-re
case.term	terminative case
cl.focus	Focus clitics
cl.quot	The quotative clitic
cv.abl	ablative converb
cv.agn	agentive case
cv.all	allative converb
cv.are	The converb -a-re
cv.ass	associative converb
cv.cont	continuing converb
cv.ela	elative converb
cv.fin	final converb
cv.gen	genitive converb
cv.imp	imperative converb
cv.impf	imperfective converb
cv.loc	locative case
cv.odd	odd converbs
cv.ques	question converb

cv.rung	The converb rung
cv.sem	semi-final converb
cv.term	terminative converb
d.dem	demonstratives
d.det	determiner
d.emph	emphasis
d.indef	indefinitie
d.plural	plurals
d.tsam	tsam and words with similar distribution
interj	interjections
n.count	count nouns
n.mass	mass nouns
n.prop	proper nouns
n.rel	relator nouns
n.v.aux	Nominalization of auxilliary verbs
n.v.cop	Nominalization of copulas
n.v.fut	Nominalization of a future verb stem
n.v.fut.n.v.past	Nominalization of a future or past stem of a verb
n.v.fut.n.v.pres	Nominalization of a future or present stem of a verb
n.v.imp	Nominalization of an imperative stem of a verb
n.v.invar	Nominalization of [v.invar]
n.v.neg	Nominalized form of the negative verb med
n.v.past	Nominalization of a past verb stem
n.v.past.n.v.pres	Nominalization of a past or present stem of a verb
n.v.pres	Nominalization of a present verb stem
neg	negation
num.card	cardinal numbers
num.ord	ordinal numbers
p.indef	indefinite pronoun
p.interrog	interrogative pronouns
p.pers	personal pronouns
p.refl	reflexive pronouns
p.refl	Reflexive pronoun
punc	punctuation
skt	Sanskrit or other metalingustic elements
v.aux	Auxiliary verbs
v.cop	Copula verbs

v.cop.neg	The negative copula verb
v.fut	Future stem of a verb
v.fut.v.past	Future or past stem of a verb
v.fut.v.pres	Future or present stem of a verb