Morphosyntactic Analyzer for the Tibetan Language: Aspects of Structural Ambiguity

Alexei Dobrov, Anastasia Dobrova, Pavel Grokhovskiy,
Nikolay Soms, and Victor Zakharov

Saint-Petersburg State University, Saint-Petersburg, Russia
LLC "AIIRE", Saint-Petersburg, Russia
a.dobrov@spbu.ru, adobrova@aiire.org, p.grokhovskiy@spbu.ru,
nsoms@aiire.org, v.zakharov@spbu.ru

Abstract. The paper deals with the development of a morphosyntactic analyzer for the Tibetan language. It aims to create a consistent formal grammatical description (formal grammar) of the Tibetan language, including all grammar levels of the language system from morphosyntax (syntactics of morphemes) to the syntax of composite sentences and supra-phrasal entities. Syntactic annotation was created on the basis of morphologically tagged corpora of Tibetan texts. The peculiarity of the annotation consists in combining both the immediate constituents structure and the dependency one. An individual (basic) grammar module of Tibetan grammatical categories, its possible values, and restrictions on their combination are created. Types of tokens and their grammatical features form the basis of the formal grammar being produced, allowing linguistic processor to build syntactic trees of various kinds. Methods of avoiding redundant structural ambiguity are proposed.

Keywords: corpus linguistics, Tibetan language, morphosyntactic analyzer, tokenization, immediate constituents, dependency grammar, natural language processing

1 Introduction

In order to build a morphosyntactic analyzer of Tibetan texts it is necessary to create a formal grammar, which includes all levels of the grammatical system of the Tibetan language from morphosyntax (syntactics of morphemes) to the syntax of sentences and supra-phrasal units.

A few currently available studies of the Tibetan language analyze mainly Tibetan morphology, the only notable exception being "The Classical Tibetan language" by Stephen Beyer [1], which also includes an extensive presentation of Tibetan syntax. Still, this work does not fully describe the Tibetan system of syntactic units and often has a speculative character, since the conclusions are not supported by textual corpora.

The current project has the following objectives:

1. To create a system of syntactical annotation of Tibetan texts, including the information about Tibetan grammatical categories, their possible values, and restrictions on their combinations;

P. Sojka et al. (Eds.): TSD 2016, LNAI 9924, pp. 1–8, 2016.
This is preprint prepared by Proceedings editor for Springer International Publishing Switzerland.

2. To develop a formal grammatical module of the open natural language processing system, which is able to perform a complete morphological and syntactic analysis of Tibetan texts;
3. To annotate a corpus of Tibetan texts syntactically.

The developed tools of language processing allow automatic markup procedures for further extension of the corpus.

The project uses an innovative approach to syntactic analysis, combining the immediate constituents structure (CS) and the dependency structure (DS). Such combination was proposed in [2] for the first time, but the available mathematical model did not allow to implement it in an algorithm. This study takes advantage of the AIIRE linguistic processor (Artificial Intelligence-based Information Retrieval Engine), which is one of the most successful computer realizations of combined CS and DS analysis [3]. Still, in order to be apllied to the Tibetan language, it requires a new research on Tibetan syntax.

2 The project's corpora resources and software

The project's database comprises two corpora of the Tibetan language developed at the Saint-Petersburg University. The Basic Corpus of the Tibetan Classical Language includes texts in a variety of classical Tibetan literary genres. The Corpus of Indigenous Tibetan Grammar Treatises consists of the most influential grammar works, the earliest of them proposedly dating back to 7th-8th centuries. Both corpora are provided with metadata and morphological annotation.

The corpora comprise 34,000 and 48,000 tokens, respectively. Tibetan texts are represented both in a Tibetan Unicode script and in a standard Latin transliteration [4].

The AIIRE linguistic processor with an open code is used for the project. AIIRE implements the method of inter-level interaction proposed by G. Tseitin in 1985 [5], which ensures the effective ambiguity resolution, based on the rules.

The principle of inter-level interaction helps to minimize the combinatorial explosion, which is very important for NLP software. The formal grammar analysis produces a considerable rate of ambiguity, especially when ellipsis is possible. The principle of inter-level interaction, implemented in the AIIRE linguistic processor, allows to apply upper-level constraints to lower-level ambiguity, and thus reduces the number of produced combinations.

The architecture of AIIRE and the developed algorithms of text analysis allow to apply this technology to languages of different types in the form of independent language modules, while the analysis algorithms are independent of the language. Besides the modules for the Russian language, modules for Arabic and Abkhaz languages were previously created, and the present project aims at developing a module for the Tibetan language, which is well known for the absence of formally marked word boundaries and ambiguity of word segmentation as such.

3 Representation of Tibetan morphological structures in AIIRE

The linguistic processor needs to recognize all the relevant linguistic units in the input text. For inflectional languages the input units are easy to identify as word forms, separated by space, punctuation marks etc. It is not the case for the Tibetan language, as there are no universal symbols to separate the input string into words or morphemes.

The developed module for the Tibetan language performs the segmentation of the input string into morphemes by using the Aho-Corasick algorithm (by Alfred V. Aho and Margaret J. Corasick), that allows to find all possible substrings of the input string according to a given dictionary. The algorithm builds a tree, describing a finite state machine with terminal nodes corresponding to completed character strings of elements (in this case, morphemes) from the input dictionary.

Language module contains a dictionary of morphemes, which allows the machine to create the tree in advance at the build stage of the language module, while in the runtime of the linguistic processor the tree is being loaded as a component of an executable module which brings its initialization time to minimum.

Two special files were created in order to analyze Tibetan morphology and morphonemics: the grammarDefines.py file determines types of tokens, their properties and restrictions, while the atoms.txt file (the allomorphs dictionary) specifies the morpheme, the token type and properties for each allomorph, also in accordance with grammarDefines.py file. For example, the following entry in the allomorphs dictionary `དགའ་|morpheme=དགའ་|type=v_root|mood=ind|has_tense=False|fin_phoneme=vowel` indicates that the དགའ་ (dga') allomorph is the basic allomorph of the དགའ་ (dga') morpheme, that is the verb root in the indicative mood, having no tense property and ending in a vowel.

The materials processed on the pilot stage allow to identify the following token types: `v_suff` (verbal suffix), `punct` (punctuation mark), `p_dem_root` (the root of the demonstrative pronoun), `n_root` (noun root), `p_pers_root` (the root of the personal pronoun), `case_marker`, `v_root` (verbal root), `num_root` (numeral root), `p_def_root` (attributive pronoun), `fin` (statement end marker). All these types of tokens have their possible morphological and morphonemic features indicated in the grammarDefines.py file. For example, the verbal root has such potential properties indicated as the mood (indicative, imperative), the tense (present, past, future), the availability of tense category (true / false) and the type of final phonemes defining the compatibility of the verbal root with suffix allomorphs. The restrictions for the verbal root require that the category of tense is available only if respective parameter "has tense" is set to "true", and the parameter of "mood" is set to "indicative".

These types of tokens and their grammatical features form the basis of the formal grammar being developed, allowing the linguistic processor to build syntactic treebanks of various structure.

Case markers of the Tibetan language, unlike inflected languages, function as postpositions rather than as suffix morphemes; and the most appropriate model to correspond to Tibetan morpheme order is seen as representing the nominal phrases followed by a case marker in postposition. Case marker takes the final position after all the modifiers of the nominal phrases, including numerals and pronouns. In this case, the order of English morphemes is opposite: the phrase dus gcig na is translated as at (na, locative) one (gcig) time (dus). Both the numeral and the nominal phrase modified by it may be further modified: the numeral may be complex, and the nominal phrase may be modified by an adjective or a participle etc. Thus, it seems to interpret the case marker of the Tibetan morphosyntax as a major constituent, and not as a dependent one, that corresponds, for example, to prepositions in prepositional groups in English and Russian languages.

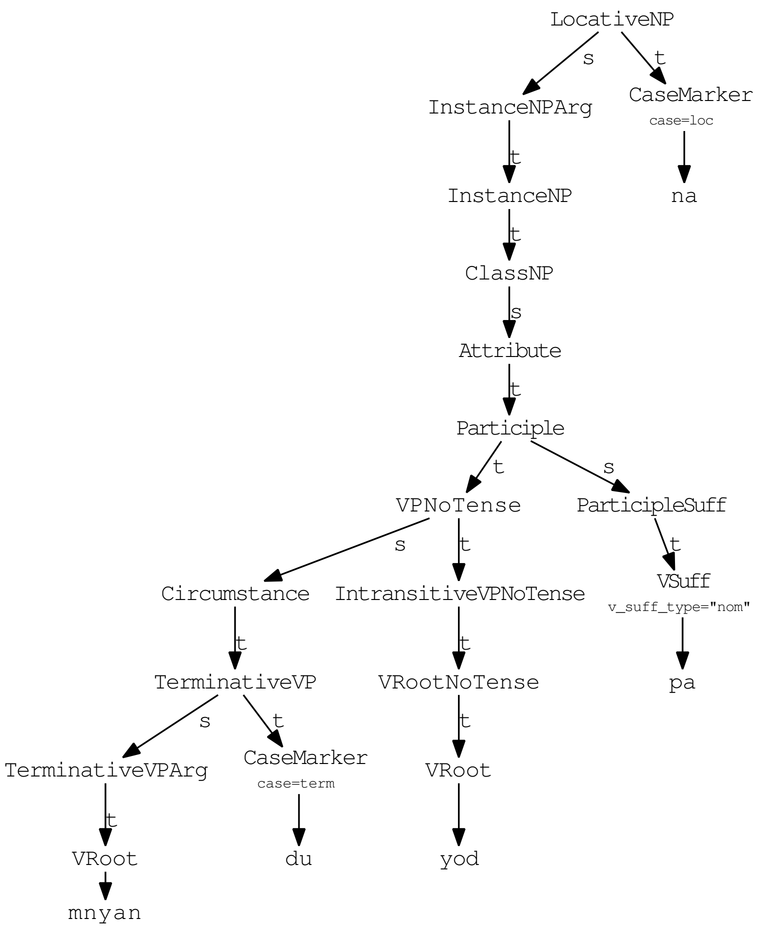

Fig. 1. Locative NP exemplified by a participle clause with a terminative adverbial modifier

Nominal phrases followed by a postpositive case marker may have structures of any complexity, including those modified by complex participle clauses, sometimes without a head, as shown in Figure 1. Such nominal phrases are often proper names or epithets (in this case it is the Tibetan morpheme-for-morpheme rendering of the Sanskrit name of the Indian city of Śravasti); the head constituent "city" being omitted due to its semantic redundancy. In this example, there is a participle verbal phrase, modified by a circumstance, which is expressed by a terminative nominal phrase (TerminativeNP), where the termenative case marker follows the masdar nominal phrase (MasdarNP), that is expressed by a verbal root (V_Root). The masdar nominal phrase omits the nominalizer, that is typical for Tibetan complex verbal nouns, including proper names: the nominalizer may be omitted both by participles and masdars, and for the time being current authors have not identified the precise rules of such omission. Literally translated, the given example reads as follows:

to hear + nominalizer omitted (missing in the tree) +
for (terminative) +
to exist + nominalizer +
in (locative)

That is, in Existing-To Be-Heard (where Existing-To Be-Heard is a name of the city).

The above mentioned features of the Tibetan morphosyntax cause a considerable rate of ambiguity in Tibetan text while being processed by a computer: due to ellipsis every verbal root can be treated as a participle or a masdar within a personal name, and each modifer can be treated as a separate proper name. As in other languages, circumstances and complements can get ambiguous interpretations if there are several recursive verbal phrases.

4 Avoiding redundant structural ambiguity: undocumented restrictions on Tibetan syntax

Ambiguity of formal syntactic structures is often produced not merely by intrinsic linguistic units' polysemy, but rather by combinatorial redundancy of the formal grammar itself. Nevertheless, exactly in these fairly frequent cases, ambiguity of formal structures shows lack of accuracy in conventional informal descriptions of language, and works as a clue to choose one of several possible ways to specify these descriptions.

As for Tibetan grammar, this on-going study has already shown some formal ambiguity cases of this kind. The examples below show how only three cases of description opacity can produce a combinatorial explosion in quite a short sentence (འདི་སྐད་བདག་གིས་ཐོས་པ་, The story about this is heard by me / The one who told this is heard by me / Those who told this are heard by me).

First off, it is not strictly specified in any of existing Tibetan grammar descriptions, including the most detailed one [1], if Tibetan predicates can be omitted. It is known that link-verbs are omitted in composite nominal predicates, but it is not clear, whether the whole predicative VP can be omitted like in Russian, or it is obligatory in any sentence like in English.

As Figure 2 shows, allowing predicate ellipsis makes the analyzed sentence about 5.3 times more ambiguous (333 vs. 63 versions of parsing). Supposing predicate ellipsis produces not only obvious versions like 'The story about this that I heard (zero predicate)', but also quite weird hypotheses like 'This (zero predicate), the story (zero predicate), (something) heard by me (zero predicate)', which seem to be ungrammatical, and a lot of their possible combinations.

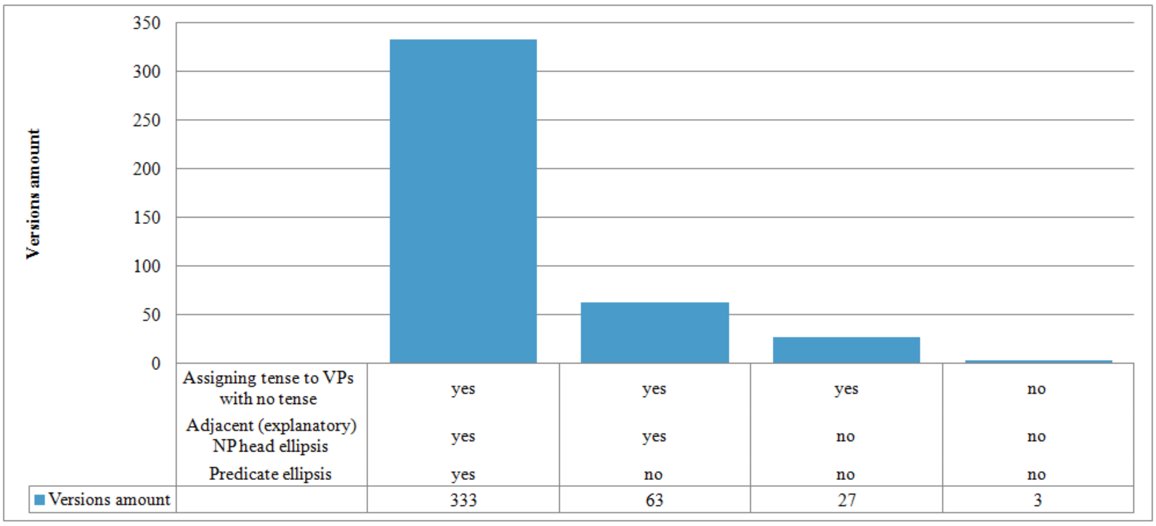

Fig. 2. Amount of versions depending on ways to formalize Tibetan syntax

Another problem of Tibetan syntax formalization that has proven to be quite important is the question about NP head ellipsis limitations. Nouns are very often omitted in Tibetan, especially in proper names, but it is absolutely unclear from existing linguistic descriptions, if such ellipsis is possible in adjacent NPs (e.g., རྒྱལ་བུ་རྒྱལ་བྱེད་ཀྱི་ཚལ་མགོན་མེད་ཟས་སྦྱིན་གྱི་ཀུན་དགའ་ར་བ་, king Making-Victory's grove - the (merchant) Giving-Food-To-the-Unprotected-Ones's amusement park). Fig. 2 shows that allowing ellipsis in adjacent NPs makes ambiguity level more than 2.3 times higher (63 vs. 27 options). This ellipsis, when allowed, produces ambiguity for each attribute of an NP, as attributes are postpositional in Tibetan and never have any markers to distinguish them from adjacent NPs. Prohibiting ellipsis in such NPs can be achieved by creating a separate constituent class for them. Generally, we can say that ellipsis and related issues in Tibetan require more theoretical research than is currently available in [1,6,7], to give a few examples.

One of the most important difficulties in Tibetan grammar formalization, however, is the problem of verbal tense. Tense is not expressed by any separate marker, but is denoted by verbal root allomorph itself. The problem is that not nearly all Tibetan verbal roots have different allomorphs for different tenses; it is the case for many verbs that tense remains unexpressed in the sentence at all. There are two options to deal with this phenomenon in formal grammar: 1) to build hypotheses for all three possible tenses for each verb root 2) to create different constituent classes (with tense feature and without tense feature) for sentences, predicates, VPs, participle and masdar phrases, etc. First option may seem attractive, as it allows to make grammar shorter, but, as Fig. 2 shows, it makes the above-mentioned sentence 9 times more ambiguous (3 versions for 2 verbs make 3*3 = 9 combinations). The conclusion is therefore obvious, that the second option, which means that such verbs and their phrases are not ambiguous in terms of tense, but rather have no tense at all, is far more plausible for Tibetan.

5 Evaluation

Analysis of 94,814 versions of parsing a sample of 374 different morpheme combinations (from 2-morphemic, and up to 36-morphemic) from the developed part of corpus has shown that all versions of parsing are correct from the formal point of view, except for 18361 versions that violate the requirements of verb-meaning dative object government model (which are pretty much semantic, and not syntactic). This result corresponds to 100% recall, and, at least, 80% precision of analysis. However, the average number of formally correct parsing versions for each combination was 974, with an exponential dependence on the number of morphemes, approximated by (1).

$$2,0403\, e^{0,2336\, x} \tag{1}$$

E.g., there are 3,020 formally correct versions of parsing, for example, for a compound sentence 'འདི་ སྐད་ བདག་ གིས་ ཐོས་ པ་ དུས་ གཅིག་ ན། བཅོམ་ ལྡན་ འདས་ མཉན་ དུ་ ཡོད་ པ་ ན་ རྒྱལ་ བུ་ རྒྱལ་ བྱེད་ ཀྱི་ ཚལ་ མགོན་ མེད་ ཟས་ སྦྱིན་ གྱི་ ཀུན་ དགའ་ ར་ བ་ ན་ བཞུགས་ སོ།།'. Analysis of these 3020 versions has shown the urgent need for different semantic restrictions, and the impossibility to resolve this ambiguity at the level of syntax.

6 Conclusion and Further Work

Computational linguists dealing with Tibetan language data face new kinds of challenges which are characteristic of this language combining isolation with agglutination. It turns out that many traditional techniques and concepts are not directly applicable for this data, and new ways of text processing should be developed.

We have identified some problems which arise during development of a morphosyntactic analyzer for Tibetan texts, and offered some solutions, such as adjacent NP head and predicate ellipsis prohibition, and discarding artificial assignment of tense to VPs with no tense.

Further prospects imply that (1) more theoretical research should be done on ellipsis in Tibetan, (2) more available corpus material should be syntactically annotated to reveal the maximum scope of syntactic constructions of Tibetan, (3) semantic dimension is to be added to the current version of Tibetan language module of the morphosyntactic analyzer in order to reduce ambiguity of Tibetan syntax further at later stages of this research.

Acknowledgment. Development of the morphosyntactic analyzer for the Tibetan language is supported by the grant No. 16-06-00578 "Morphosyntactic Analyser of Texts in the Tibetan language" by Russian Foundation for Basic Research.

Authors also acknowledge Saint-Petersburg State University for a research grant 2.38.293.2014 "Modernizing the Tibetan Literary Tradition" which enabled the conceptual study of the original Tibetan texts.

References

1. Beyer, S.: The Classical Tibetan language. State University of New York, New York (1992)
2. Gladkii, A.V.: Syntactic structures of natural language in automated communication systems [Sintaksicheskie struktury estestvennogo jazyka v avtomatizirovannyh sistemah obshhenija]. Nauka, Moscow (1985)
3. Dobrov, A.V.: Automatic Classification of News by Means of Syntactic Semantics [Avtomaticheskaja rubrikacija novostnyh soobshhenij sredstvami sintaksicheskoj semantiki], Doctoral Thesis. Saint-Petersburg State University (2014)
4. Grokhovskiy, P., Khokhlova, M., Smirnova, M., Zakharov, V. Tibetan Linguistic Terminology on the Base of the Tibetan Traditional Grammar Treatises Corpus. In: Lecture Notes in Computer Science, Vol. 9302, 2015, pp. 326–333.
5. Tseitin, G.S.: Programming in Associative Networks [Programmirovanie na associativnyh setjah], Computers in designing and manufacturing [EVM v proektirovanii i proizvodstve] (2). Mashinostroenie, Leningrad, pp. 16-48 (1985).
6. Andersen, Paul Kent.: Zero-anaphora and related phenomena in Classical Tibetan. In: Studies in Language, Vol. 11, pp. 279–312.
7. Denwood, P.: Tibetan. Amsterdam, John Benjamins Publishing, 1999.