Tibetan Nominalized Verb Phrases and Their Modeling in the Formal Grammar and the Computer Ontology

A. V. Dobrov, A. E. Dobrova, O. V. Dzhangolskaya, M. P.Smirnova, and N. L. Soms

***Abstract*—The Tibetan language is characterized by widespread of nominalizing particles and noun-nominalizers as well as different types of nominalization constructions (clausal nominalization, action nominalization). Regular nominalizers (i.e., nominalizing particles) convert a verb or a verb phrase into a semantically neutral proposition that can be used in a nominal context. Noun-nominalizers can also function as nouns and usually occur in ambiguous context. Such nominalizers add specific meaning to the newly formed proposition. The phenomenon of nominalization is still subject to thorough investigation based on corpus data. Given article describes different types of nominalizers, types of constructions with them, and methods of their modeling in the formal grammar and the computer ontology.**

***Keywords*—Computer ontology, nominalization, Tibetan language, Tibetan verbs.**

I. Introduction

The research introduced in this paper is focused on the study of nominalization in the Tibetan language and developing methods for nominalized verb phrases modeling in the formal grammar and the computer ontology. This research is a part the work on development methods of Tibetan NLP aimed at the creation of a full formal model of the Tibetan language that will allow the linguistic processor to perform a correct annotation (including morpho-syntactic, syntactic, and semantic parsing) of Tibetan texts. The development of a sufficient formal model requires creation of a corpus of texts with a decent annotation. For its part, the creation of a reliable annotation calls for a formal model to serve as a basis.

The present formal model of a Tibetan grammar, vocabulary and ontology were created with the support of a corpus of texts containing Tibetan grammatical texts and texts on the theory of writing in Classical and Modern Tibetan (69388 tokens). The corpus is provided with metadata and morphological annotation.

One of the current problems is methods of formal grammatical and ontological modeling of Tibetan nominalized verb phrases. Nominalization in the Tibetan language is a quite frequently used linguistic phenomenon. Nominalizers transform a verbal proposition of any length and complexity in a nominalized verb phrase that can occur in a sentence in a syntactically nominal context. To date, there is no systematic description of Tibetan nominalizers in Tibetological literature, confirmed by corpus data. For this reason, their modeling in the formal grammar and the computer ontology is connected with a number of difficulties.

First of all, there are different types of nominalizers (real and noun-nominalizers) in the Tibetan language that function in a diverse way. Nominalizers occur in various types of constructions, and each of them requires a separate approach in modeling. The context often does not provide enough information to conclude whether it is a nominalizer, a noun or a part of a compound. Moreover nominalized verbs and verbal phrases can be idiomatized and thus require a specific modelling in the computer ontology.

II. Related Work

Nominalization is a conversion of a verbal construction of any length and intricacy into a newly-formed proposition that can occur in a regular nominal context anywhere in a sentence [1, p. 294]. This grammatical phenomenon is typical to many Tibeto-Burman languages where it can also function as a derivation instrument, forming new lexical nouns or adjectives [2, p. 163].

In some cases the nominalizer doesn’t make the whole clause nominalized, but only the verb. The verb dependents are treated as nominal dependents and do not require the accordance with verb transitivity. This type is quite typical for the Tibeto-Burman languages and is called an “action nominalization” or “event nominalization” [2, p. 166]. In the texts of our corpus, we found both – cases of standard clausal nominalization and action nominalization. The latest, however, were not specifically described for the Tibetan language. Nonetheless this phenomenon causes a number of difficulties for formal grammatical and

Manuscript received August 25, 2020. This work was supported by the Russian Foundation for Basic Research, Grant No. 19-012-00616, “Semantic Interpreter of the Texts in the Tibetan Language.”

A. V. Dobrov is with Saint Petersburg State University, Russian Federation (e-mail: a.dobrov@spbu.ru).

A. E. Dobrova is with the LLC “AIIRE”, Saint Petersburg, Russian Federation (e-mail: adobrova@aiire.org).

O. V. Dzhangolskaya is with Saint Petersburg State University, Russian Federation (e-mail: lamenth@yandex.ru).

M. P.Smirnova is with Saint Petersburg State University, Russian Federation (phone: +7 911-127-7186; e-mail: m.o.smirnova@spbu.ru).

N. L. Soms is with the LLC “AIIRE”, Saint Petersburg, Russian Federation (e-mail: nsoms@aiire.org).

ontological modeling that will be described below.

Nominalizers of the Tibeto-Burman languages can bear more than just a nominalizing function. They can convey an additional specific meaning (e.g., place of action) [2, p. 170].

Depending on the meaning of nominalizer and the meaning of the formed nominalized verb phrase S. Beyer divides Tibetan nominalizers into two categories: patient-centered and proposition-centered nominalizers. Patient-centered nominalizers convey the meaning of a certain aspect of a patient of a proposition nominalized. S. Beyer gives three types of patient-centered nominalizers. The nominalizer *-rgyu* denotes 'patient of proposition' like in (1).

(1) དཔྱད་པ་གཏོང་རྒྱུ
dpyad-pa gtong rgyu
analyse-NMLZ abandon-NMLZ
'a cause to abandon the analysis'

Nominalizers *-'o-cog/-dgu/-tshad* denote all patients of proposition like in (2).

(2) བུ་དང་བུ་མོ་བཙའོ་ཅོག
bu dang bu-mo btsa'o-cog
son CONJ daughter
bear-NMLZ
'all the sons and daughters [she] bears'

Nominalizers -*'phro/'phros* denote 'remainder of patient of proposition' like in (3) [1, p. 296-298].

(3) ཡི་གེ་འབྲིས་འཕྲོས
yi-ge 'bri-'phros
letter write-NMLZ
'part of a letter that [someone] is writing'

Second type of nominalizers, indicated by S. Beyer, conveys the meaning of an entire proposition nominalized. This type includes the following particles: *-Pa*, *-sa*, *-grogs*, *-mkhan/-mi*, *-tshul*, *-nyen*, *-dus*, *-res*, *-lugs*, *-thabs*, *-grabs* [1, p. 295]. Some of them are considered by S. Beyer to be real nominalizers (like *-sa* 'place,' *-grogs* 'help,' and -*mkhan/-mi* 'person') and some - quasi-nominalizers (*-tshul* 'way,' *-nyen* 'danger,' *-dus* 'time,' *-res* 'turn at,' *-lugs* 'method,' *-thabs* 'opportunity,' *-grabs* 'preparation'). Quasi-nominalizers can be interpreted as nouns that are slowly turning into nominalizing particles [1, p. 294]. S. Beyer mentions that quasi-nominalizers can be used not only as nominalizers after a verb, like in (4), but also in noun phrases with genitive after nominalized verb like in (5).

(4) དམྱལ་བར་འགྲོ་ཉེན
dmyal-bar 'gro-nyen
hell-LOC go-NMLZ
'danger of going to hell'

(5) འགྲོ་བའི་གྲབས
'gro-ba 'i grabs
go-NMLZ GEN preparations
'preparations to leave'

It should be noted that the status of all proposition-centered nominalizers except *-Pa*, which is the most common nominalizer, cannot be considered unambiguous, as they also function as nouns or parts of compounds and their meanings as nominalizers derive from their meanings as nouns.

III. The Software Tools for Parsing and Formal Grammar Modeling

This study was performed with use of and within the framework of the AIIRE project [3]. AIIRE is a free open-source NLU system, which is developed and distributed in terms of GNU General Public License (http://svn.aiire.org/repos/tproc/trunk/t/).

This framework implements the full-scale procedure of natural language processing, beginning from graphematics (Aho-Corasick algorithm had to be used for the Tibetan language due to absence of word delimiters), continuing with morphological annotation, going further with syntactic parsing, and ending with semantic analysis.

The morphemic dictionaries developed for the morphological annotation for the Tibetan Language were described in the article [4] and are not relevant to this paper.

Syntactic parsing is performed in terms of a combined constituency and dependency grammar, which consists of the so-called classes of immediate constituents (hereinafter CICs). These classes are developed as python-classes, with the builtin inheritance mechanism involved, and provide attributes that specify the following information:

1) The template of semantic graph which represents the meaning of this constituent;
2) The list of possible head constituent classes;
3) The list of possible subordinate constituent classes;
4) The dictionary of possible linear orders of the subordinate constituent in relation to the head and the meanings of each order;
5) The boolean field for head ellipsis possibility;
6) The boolean field for subordinate constituent ellipsis possibility;
7) The boolean field for possibility of non-idiomatic semantic interpretation.

Due to the absence of word delimiters and any formal evidence of boundaries between morphology and syntax, Tibetan texts have to be parsed by morphemes instead of being parsed by wordforms, as it can be done for Indo-European languages. Therefore, the formal grammar contains CICs both for regular syntactic models and for models which are usually treated as word-formational, in particular some models of derivates (there only a few of them) and models of compounds.

The grammar is developed in straight accordance with semantics, in a way that the meanings of syntactic and morphosyntactic constituents can be correctly evaluated in accordance with the Compositionality principle. Each constituent is provided with a set of semantic interpretations on the stage of the semantic analysis; if this set proves to be empty for some versions of constituents, then these versions are discarded; this is how syntactic disambiguation is performed. The results of semantic analysis are stored as semantic graphs, but, for idioms like compounds, these graphs consist of single concepts, thus, the structure of semantic graphs is not a matter of discussion in this article.

IV. THE SOFTWARE TOOLS FOR ONTOLOGICAL MODELING

The ontology is implemented within the framework of AIIRE ontology editor software; this software is free and open-source, it is distributed under the terms of GNU General Public License [9; 10], and the ontology itself is available as the snapshot at [11] and it is also available for unauthorized view or even for edit at [12] (edit permissions can be obtained by access request).

The ontology, used for this research, is a united consistent classification of concepts behind the meanings of Tibetan linguistic units, including morphemes and idiomatic morphemic complexes. The concepts are interconnected with different semantic relations. These relations allow to perform semantic analysis of texts and lexical and syntactic disambiguation. The basic ontological editor is described with examples from the Tibetan ontology in articles [4], [5] and [6].

Modeling verb meanings in the ontology is associated with a number of difficulties. First of all, the classification of concepts denoted by verbs should be made in accordance with several classification attributes in the same time, which arise primarily due to the structure of the corresponding classes of situations that determine the semantic valencies of these verbs. These classification attributes are, in addition to the semantic properties themselves (such as dynamic / static process), the semantic classes of all potential actants and circumstants, each of which represents an independent classification attribute. With the simultaneous operation of several classification attributes, the ontology requires classes for all possible combinations of these attributes and their values in the general class hierarchy. Special tools were created to speed up and partly automate verbal concepts modeling. AIIRE Ontohelper is used together with the main AIIRE ontology editor web interface to build the whole hierarchy of superclasses for any verb meaning in the ontology. The structure and operation of the Ontohelper editor are described in detail in [7, p. 147].

V. NOMINALIZERS IN THE TIBETAN CORPUS

1. Real Nominalizers

The study of the corpus allowed us to identify only two real nominalizers, i.e. particles that perform only nominalizing function, have standard set of common (neutral) meanings, and don't occur in ambiguous context. The most general and frequently used nominalizer is *-Pa*. It has allomorphs depending on the preceding final: *-pa* after *-g, -d, -n, -b, -m*, and *-s*; *-ba* after preceding *-r*, *-l*, and open syllables. For it the CIC NominalizerSuff was created in the formal grammar. This class is embedded as a modifier in the classes for nominalized verbal phrases of different types (VNNoMorphon, VNNoMorphonEllArg, VNNoTenseNoMorphon, VNNoTenseNoMorphonEllArg, VNNoTenseNoMoodNoMorphon, VNNoTenseNoMoodNoMorphonEllArg).

The nominalizer is commonly used in a neutral context. It signals only that the verb or entire proposition is functioning as a nominal, and contributes nothing further to the meaning of a newly formed proposition [1, p. 295]. The verbal phrase with *-Pa* can denote a process, a subject, and an object of an action (for transitive verbs). Thus, one nominalized verbal phrase usually has two (for intransitive verbs, e.g., (6)) or three (for transitive verbs, e.g., (7)) semantic versions of parsing.

(6) གཤེགས་པ
gshegs-pa
come-NMLZ
(6.1) 'coming'
(6.2) 'one who came'

(7) བསླབ་པ
bslab-pa
teach-NMLZ
(7.1) '[the process of] teaching'
(7.2) 'one who taught'
(7.3) 'teaching [that is taught]'

The graphs of semantic parsing for (7.1-3) are represented on the Fig. 1.

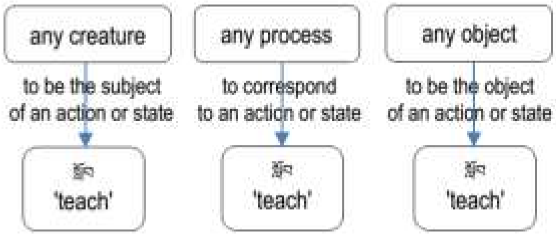
any creature
to be the subject of an action or state
སློབ
'teach'
any process
to correspond to an action or state
སློབ
'teach'
any object
to be the object of an action or state
སློབ
'teach'

Fig. 1. Semantic graphs for the nominalized verb *bslab-pa*

The second nominalizer that we consider to be real is *-rgyu*. Despite the fact that it is also used in the corpus texts as a noun with the meaning 'reason' and as part of compounds, in all cases its meaning and function is unambiguous. For *-rgyu* and for other nominalizers that do not have allomorphs (if such will be discovered in the future), a common class NominalizerSuffNoMorphon was created in the formal grammar.

Real nominalizers usually function as standard clausal nominalizers (nominalize the whole verbal phrase). Cases of action nominalization also occur (for examples see section 3.4 of this article), however, far less frequently. For example, in one of the texts of the corpus with a volume of 13462 tokens there are 191 cases of nominalization with *-Pa*. However, only five of them are considered to be action nominalization.

2. Zero Nominalization

The term zero nominalization was suggested by N. Hill for morphologically finite forms occurring in syntactically nominal contexts [8, p. 5]. S. Beyer describes similar cases when the nominalizer *-Pa* can be omitted between a tense stem of a verb and a bound role particle [1, p. 305].

Examples of this phenomenon can be found in the corpus poetic texts. In most of them the right context indicates that a verb functions as a noun. Usually the nominalizer *-Pa* occurs only after the second of two verbs while the nominalization of the first is guaranteed by the choice of conjunction particle *dang* like in (8), since *dang* occurs only after nouns or noun phrases [8, p. 5; 1, p. 241].

(8) དགར་དང་བརྣན་པའི་ཚིག་ཏུ་འགྱུར
dgar dang brnan-pa 'i tshig tu 'gyur
separate CONJ emphasise-NMLZ GEN phrase TERM
become

'[it] becomes the term of segregation and stress'

In example (9) we meet three cases of the nominalizer -*Pa* omission after the verbs *'dri* 'to ask,' *klog* 'to read,' and *bshad* 'to speak.'

(9) འདྲི་དང་ཀློག་དང་བཤད་རྣམས་ཀྱི། །མཚམས་སྦྱོར་སྒྲ་ལ་ཐོགས་མེད
'dri dang klog dang bshad rnams kyi/ /mtshams-sbyor sgra la thogs med
ask CONJ read CONJ speak-PL GEN conjoining_marker DAT obstruct not_exist
'there will be no difficulties with markers linking [words in the process of] writing, reading and explaining'

In the first two cases of zero nominalization in (9) the choice of conjunction particle *dang* guarantees the interpretation of *'dri* and *klog* as nominal forms. After *bshad* we meet the plural marker *rnams* that also follows only nouns or noun phrases. For such cases in the formal grammar the CIC poetic verbal noun (PoeticVN) was created. This class was embedded in the CIC for noun phrases in plural (InstanceNPPlural) and homogeneous noun phrases (InstanceNPGroup).

Few examples of the same phenomenon were found in the prose texts of the corpus. However, to claim that zero nominalization is typical not only for poetry and can be considered as an ellipse, valid in the whole language, it is necessary to determine more typical grammatical contexts of zero nominalized forms. Otherwise, the development of the formal grammar in such a way will multiply the number of verbal noun classes, which could potentially lead to combinatorial explosions.

Unlike the two previous examples, where the right context after each verb makes it possible to interpret them as nominalized forms, more difficult cases can be found when some noun coordinators are used once at the end of the passage like in (10) and (11).

(10) རྗེས་འཇུག་བཅུ་ཡི་སྦྱོར་བ་ནི། །མཉན་བསམ་བསྟན་པའི་དོན་དུ་སྦྱར། །
rjes-'jug bcu yi sbyor-ba ni/ /mnyan bsam bstan-pa'i don du sbyar/ /
final_consonant ten GEN join-NMLZ TOP listen think teach-NMLZ
'As for adding of the ten final consonants, [these consonants] are added for listening, thinking and teaching.'

In example (10) the nominalizer -*Pa* is used once after three verbs – *mnyan* 'to listen,' *bsam* 'to think,' and *bstan* 'to teach,' that can be considered as homogeneous verbal phase. As it is not a typical grammatical phenomenon for the Tibetan language the special class PoeticHomogenVP, that was embedded into classes for verbal nominalization.

(11) སྡེབ་སྦྱོར་ལེགས་མཛད་མཁས་རྣམས
sdeb-sbyor legs mdzad mkhas rnams
poetry be_good do be_skilled-PL
'[those who are] skilled in making good poetry'

In example (11) we actually see five verbs with obviously different subordinate syntactic relations but without any grammatical markers between them. Only the last verb takes the plural marker and thus can be undoubtedly treated as a case of zero nominalization. Still this passage can be read in several ways. To perform disambiguation in this case several combinations of verbs are treated as compounds of different types.

3. Noun-nominalizers

The term "noun-nominalizer" (i.e., quasi nominalizers) means nouns that are adjacent to the stem of the verb directly and function as nominalizers. Noun-nominalizers give additional meaning to the nominalized verb or proposition. We discovered the following noun-nominalizers in the Tibetan corpus: *mkhan* 'person,' *tshul* 'way, method,' *thabs* 'skill, technique,' *stangs* 'manner,' *rtsal* 'capacity,' *lugs* 'manner,' *cha* 'part,' *sa* 'place,' and *dus* 'time.'

All noun-nominalizers can occur in the corpus as independent nouns, as nominal elements of compounds, as standard clausal nominalizers, and as action nominalizers.

3.1 Noun-nominalizers as Nouns

A noun-nominalizer used as an independent noun does not require special modeling in the ontology. It belongs to the class NRoot and functions as a regular noun. Quite often it is used after the verb that has been already nominalized with -*Pa* and put in the genitive case, as in the example (12).

(12) གཞི་གང་་དང་འབྲེལ་བའི་ཚུལ
gzhi gang dang 'brel ba'i tshul
basis INDF ASS connect-NMLZ GEN way
'way of connecting with any basis'

Such examples confirm inconstant status of these nouns as nominalizers.

3.2 Noun-nominalizers as Nominal Components of Compounds

Noun-nominalizers regularly function as parts of compounds of different types. Tibetan compounds are idiomatized contractions of two or more syntactic groups with the fixed syntactic relation inside the compound. Mostly, the grammatical morphemes that indicate these relations are omitted [1, p. 102]. For example, the noun *sa* 'place, earth', that can act as a nominalizer, occurs as the second element of almost all types of nominal compounds: compound noun root group (CompoundNRootGroup, e.g. (13)); adjunct compound (AdjunctCompound, e.g., (14)); noun phrase with genitive compound (NPGenCompound, e.g., (15)); and compound class noun phrase (CompoundClassNP, e.g., (16)).

(13) གནམ་ས
gnam-sa
heaven_earth
'heavens and earth'

(14) ས་གནས
sa-gnas
place_place
'territory'

(15) ལྷ་ས
lha-sa
god_earth
'earth of gods' (Lhasa)

(16) ས་ཆེན
sa-chen
place_be_big
'high level'

We developed specific ways of modeling the meanings of each type of compounds (see [7] for detailed classification of Tibetan compounds and ways of their modeling in the formal grammar and computer ontology). However, if the first component of NPGenCompound is a verbal root (like in (17)), sometimes such cases are difficult to separate from nominalization with *sa* like in (18).

(17) རྒྱལ་ས
rgyal-sa
obtain_victory_place
'capital (lit. 'place of the one who obtained the victory')'

(18) སྡོད་ས
sdod sa
live_place
'place of living'

Examples (17) and (18) are both idioms but modelled differently. It is necessary to establish subclass of the general genitive relation 'to have an object or process' between the components of the compound (17) in the computer ontology. Cases like (18) are treated as derivational nominalization, described in section VII of this article.

3.3 Noun-nominalizers as Clausal Nominalizers

According to the data of our corpus, unlike nominalizing particles (i.e., real nominalizers) most noun-nominalizers occur as standard clausal nominalizers and as action nominalizers with comparable frequency.

In the case of clausal nominalization, the transitivity of the nominalized verb is preserved, and the whole verbal phrase is nominalized, as in (19). In this example the noun-nominalizer *thabs* 'a skill' is used to nominalize the verb *bzo* 'to make' with its direct object *shog* 'paper'.

(19) ཤོག་བུ་བཟོ་ཐབས
shog bu bzo thabs
paper make skill
'skill of making paper'

Since noun-nominalizers can act as clausal nominalizers, while maintaining their meaning, the special property 'nominalize_verb' for noun roots was created in the formal grammar file that determines types of tokens, their properties and restrictions. For all discovered nouns-nominalizers the value of this property was set to "true".

The CIC NominalizerNRoot was added in the formal grammar with noun roots that require this property being the head class and intersyllabic delimiter being the subordinate constituent. This class is embedded as a modifier in the same classes for nominalized verbal phrases as real nominalizers.

In the computer ontology we created the same basic class for all nouns that occur as nominalizers in our corpus – 'typical agent, circumstance or mode of action or state.' At the same time, these nouns retain their unique hypernyms, which allow them to act as independent nouns, i.e. participate in different semantic relations.

The created class was connected with a specific relation 'to be an agent, circumstance or mode of action or state' with the basic class for all verbs in the ontology 'to perform an action or state' so that meanings of the nouns were reflected in semantic versions of parsing.

In the result we obtained the presented on the Fig. 2 semantic graph for parsing an example (19).

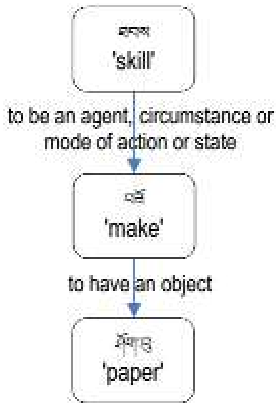

Fig. 2. Semantic graphs for the nominalized verb phrase *shog bu bzo thabs* 'skill of making paper'

3.4 Noun-nominalizers as Action Nominalizers

When noun-nominalizers are used as action nominalizers the verb loses its transitivity and the object of this verbal phrase is treated as the noun subordinate. It should be noted that almost all noun nominalizers in our corpus can be used as clausal and action nominalizers in the same texts and even in the same phrases.

In phrases with action nominalization the nominalized verb is connected with its direct object with the genitive case marker. In the example (20) noun-nominalizer *thabs* nominalizes the same verb as in (19), but the transitivity of the verb is lost.

(20) མཚལ་དང་སྣག་ཚའི་བཟོ་ཐབས
mtshal dang snag-tsha 'i bzo thabs
vermilion CONJ ink make-NMLZ
'skill of making vermilion and ink'

Real nominalizers can also be found in phrases with action nominalization, as in (21). However, such cases are much less frequent. In the example (21) the verb *sbyor* 'to join' is nominalized with *-Pa*, the most common nominalizer, but its direct object is put in the genitive case, so *sbyor-ba* is grammatically treated as a noun.

(21) ཡི་གེའི་སྦྱོར་བ་བཤད
yi ge 'i sbyor ba bshad
letter GEN join-NMLZ explain
'explain the joining of letters'

Such cases cause difficulties in semantic modeling. If we consider (20) and (21) to be cases of nominalization than in the computer ontology we should "allow" vermilion and ink from (20) and letters from (21) to have processes, that is to connect the concepts 'any group' (semantic class for the group of homogeneous nouns) and 'language unit' (the basic class of the concept *yi-ge* 'letter') with 'any process' (i.e., the basic class for all verb meanings in the computer ontology) with genitive relation, that can cause semantic ambiguity. However, this approach requires a lot of effort from an ontology editor and may cause semantic ambiguity.

In order to achieve systematic triggering of genitive relations between noun phrase and verb nominalization phrase as relations with a subject or an object in such constructions, we build a separate hierarchy of process classes, completely parallel to the hierarchy of classes of verb meanings, in the ontology. These technical classes are provided with technical names like 'process for <verb>' (e.g., 'process for སྦྱར (join...)'. Tibetan nominalizations could not be used themselves as these names, because they mean not only processes, but also attributes. For each verb meaning, a technical concept is constructed that corresponds to the process of performing the action or state that is denoted by this verb meaning. Classes of processes are provided with special relations with classes of objects and subjects, fully reproducing these relations of verb classes. These relations are built into the classification of relations as genitive relations. Thus, processes denoted by nominalizations receive valencies in a genitive construction that reproduce subject and object valencies of verbs. This solution makes it possible not to create special classes of immediate constituents for these constructions; at the level of parsing, one version is built, which excludes the possibility of technical ambiguity.

VI. Idiomaticity of Nominalized Verbs and Verbal Phrases

Nominalized verbs or verbal phrases can be idiomatized. Cases of nominalized verb idiomatization usually correspond to derivational nominalization (derivation of lexical nouns) like (22) and (23).

(22) འབྲེལ་བ
'brel-ba
join-NMLZ
'coherent speech'

(23) མཁས་པ
mkhas-pa
be_learned-NMLZ
'sage'

To ensure the correct semantic parsing of such idioms we model its meaning in the computer ontology. In addition, separate concepts must be created for all possible nominalization meanings in the ontology so that the possibility of literal interpretation is not automatically excluded. Thus, in the computer ontology additional concept with the meaning 'joining' is created for the expression (22), and concepts with the meanings 'one who knows' and 'knowing' are created for the expression (23).

Verbs and verbal phrases formed by noun-nominalizers can also be idiomatized. Such cases were also discovered in the corpus. For example, (24) and (25) are grammatical terms formed by nominalization of two transitive verbal phrases.

(24) ལ་དོན་སྦྱོར་ཚུལ
la-don sbyor-tshul
la_meaning add-NMLZ
'rules of the use of particles with the meaning of *la*'

(25) བྱེད་སྒྲ་སྦྱོར་ཚུལ
byed-sgra sbyor-tshul
do_marker add-NMLZ
'rules of the use of agent marker'

If an idiom is represented by a single verb and a noun-nominalizer (e.g., (26)) it is modeled as a compound.

(26) ཀློག་ཚུལ
klog-tshul
read_way
'transcription'

Thus, *klog-tshul* (26) obtains two versions of syntactic parsing - as nominalized verbal phrase and as noun phrase with genitive compound. This preserves the possibility of literal interpretation without special modeling in the ontology.

VII. Current Statistics

The statistics of noun-nominalizers use in the corpus is presented in the Table 1.

Table 1. Statistics on noun-nominalizers use in the current corpus

	Noun	Part of a compound	Action nominalizer	Standard clausal nominalizer	No grammatical context	Total amount
mkhan	0	8	0	12	7	27
tshul	73	11	72	53	71	280
lugs	19	58	5	13	4	99
thabs	15	6	4	9	10	44
stangs	0	0	5	11	1	17
sa	19	42	0	5	6	72
cha	45	96	3	1	5	150
rtsal	5	13	0	5	4	27
dus	96	53	0	14	11	174

As we see from Tab. 1 only *sa, mkhan, dus* and *rtsal* does not occur after verbal roots preceded by another noun in the genitive case (when the verb valency is not preserved). However there are some cases when there is no left context or the context cannot help to define whether it is a nominalizer or a part of compound. Other nouns including the most frequent can be used as head nouns in the genitive noun phrases with verbs nominalized by *-Pa*, as standard clausal and action nominalizers.

VIII. Conclusion

Most nouns that can function as standard clausal nominalizers can be considered quasi-nominalizers (in a broader sense than it was proposed by S. Beyer), as they are frequently used in alternative grammatical context: as parts of compounds, as nouns (in particular as head nouns in the genitive noun phrases with verbs nominalized by *-Pa*) and as action nominalizers after verbal roots without preserving the original verb valency. Additional complexity is created by the frequent idiomatization of nominalized verbs and verb phrases that requires modeling of their literal and idiomatic meanings in the ontology. This way of modelling leads to morpho-syntactic and semantic ambiguity. At the moment this ambiguity cannot be fully resolved, but we expect that further work and enlargement of the corpus will allow us to address this issue.

References

[1] S. Beyer, *The Classical Tibetan Language*. New York: State University of New York, 1992.

[2] C. Genetti, “Nominalization in Tibeto-Burman languages of the Himalayan area: A typological perspective,” *Nominalization in Asian Languages: Diachronic and Typological Perspectives*, 2011, pp. 163–194.

[3] A. Dobrov, A. Dobrova, P. Grokhovskiy, N. Soms, V. Zakharov, “Morphosyntactic analyzer for the Tibetan language: aspects of structural ambiguity,” presented at International Conference on Text, Speech, and Dialogue, 2016, pp. 215-222.

[4] A. Dobrov, A. Dobrova, P. Grokhovskiy, N. Soms, “Morphosyntactic Parser and Textual Corpora: Processing Uncommon Phenomena of Tibetan Language”, in *Proc. of the International Conference IMS*, pp. 143–153, 2017.

[5] A. Dobrov, A. Dobrova, P. Grokhovskiy, M. Smirnova, N. Soms, “Computer Ontology of Tibetan for Morphosyntactic Disambiguation,” in *Proc. of DTGS. Communications in Computer and Information Science*, vol. 859, pp. 336–349, 2018.

[6] A. Dobrov, A. Dobrova, P. Grokhovskiy, M. Smirnova, N. Soms, “Idioms Modeling in a Computer Ontology as a Morphosyntactic Disambiguation Strategy,” *TSD 2018. Lecture Notes in Computer Science*, vol. 11107, pp. 76–83, 2018.

[7] A. Dobrov, A. Dobrova, M. Smirnova, N. Soms, “Formal grammatical and ontological modeling of corpus data on Tibetan compounds,” in *Proc. of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Vol. 2. SciTePress, pp. 144–153, 2019.

[8] N. W. Hill, “Tibetan zero nominalization,” *Revue d’Etudes Tibétaines*, no. 48, pp. 5–9, 2019.

[9] AIIRE Ontology. Available: http://svn.aiire.org/repos/ontology/

[10] AIIRE Ontohelper. Available: http://svn.aiire.org/repos/ontohelper/

[11] Tibetan ontology. Available: http://svn.aiire.org/repos/tibet/trunk/aiire/lang/ontology/concepts.xml

[12] Tibetan ontology (unauthorized view). Available: http://ontotibet.aiire.org