Ontological and Formal Grammatical Modeling of Tibetan Nominalized Verb Phrases

M.O. Smirnova[1], A.V. Dobrov [1], A.E. Dobrova[2], N.L. Soms[2], O.V. Dzhangolskaya[1]

[1] St. Petersburg State University, [2] LLC «AIIRE»

m.o.smirnova@spbu.ru, a.dobrov@spbu.ru, adobrova@aiire.org, nsoms@aiire.org, lamenth@yandex.ru

Abstract

Nominalization in the Tibetan language is widely used: nominalizers added to verb roots can form verb phrases of any length and complexity that can be used in a syntactically nominal context. The most frequently observed Tibetan nominalizer is a nominalizing particle -Pa that is used to convert an entire preposition into a semantically neutral nominalized verb. Scholars also identify quasi-nominalizers in the Tibetan language – nouns that can be used not only as a noun but also as a nominalizing particle with specific meaning (e.g., tshul «way of»). Grammatical context does not always make it possible to define whether the word functions as a nominalizer or a noun. With a certain frequency nominalized verb phrases are idiomatized thus requiring a special way of modeling in the computer ontology. Given article describes Tibetan nominalizers and quasi-nominalizers, nominalization models and methods of ontological and formal grammatical modeling of Tibetan nominalized verbs and verb phrases.

Keywords: Tibetan Language, Compounds, Computer Ontology, Nominalization.

Reference for citation: Smirnova M.O., Dobrov A.V., Dobrova A.E., Soms N.L., Dzhangolskaya O.V. Ontological and Formal Grammatical Modeling of Tibetan Nominalized Verb Phrases // Computer Linguistics and Computing Ontologies. Vol. 4 (Proceedings of the XXIII International Joint Scientific Conference «Internet and Modern Society», IMS-2020, St. Petersburg, June 17-20, 2020). - St. Petersburg: ITMO University, 2020. P. 63 – 72. DOI: 10.17586/0000-0000-2020-4-63-72

Introduction

The research introduced in this paper was conducted within the framework of an on-going series of research projects, aimed at the development 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) without any manual corrections. 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 from The Basic Corpus of the Tibetan Classical Language, The Corpus of Indigenous Tibetan Grammar Treatises and the Corpus of Modern Tibetan literature. All texts were annotated and approved manually and displayed both in the Tibetan Unicode script and in standard Latin (Wylie) transliteration.

One of the current problems is methods of formal grammatical and ontological modeling of nominalized verb phrases. Nominalization in the Tibetan language is 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, since it is often unclear whether we should interpret them as nouns (independent or parts of compounds) or as nominalizing particles. Moreover nominalized verbs in the Tibetan language are often idiomatized – acquire a specific meaning. Such verb phrases also require a certain way of modelling in the computer ontology.

1. 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]. Depending on the language, nominalizers can bear more than just a nominalizing function, they can convey an additional specific meaning (e.g., place of action) or function as converb forms [2, p. 170].

In some cases the nominalizer doesn't make the whole clause nominalized, it nominalizes only the verb. The dependents of the verb no longer require the accordance with verb transitivity and are treated as nominal dependents. This type quite often occurs in Tibeto-Burman languages and is called an «action nominalization» or «event nominalization» [2, p. 166]. In the texts of our corpus, we found cases of nominalization of a similar type, which, however, were not specifically described for the Tibetan language. Nonetheless this phenomenon causes a number of difficulties for formal grammatical and ontological modeling that will be described below.

According to S. Beyer, nominalizers in the Tibetan language can be divided 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. This type is unusual to the speakers of English, but is natural and often used in Tibetan. S. Beyer gives three patient-centered nominalizers: *-rgyu* (denotes 'patient of proposition' like in (1)), *-'o-cog/-dgu/-tshad* (denote all patients of proposition like in (2)) and *-'phro/'phros* (denote 'remainder of patient of proposition' like in (3)) [1, p. 296-298].

(1) དཔྱད་པ་གཏོང་རྒྱུ	(2) བུ་དང་བུ་མོ་བཙའོ་ཅོག།	(3) ཡི་གེ་འབྲིས་འཕྲོས
dpyad-pa gtong rgyu	*bu dang bu-mo btsa'o-cog*	*yi-ge 'bri-'phros*
analyse-NMLZ abandon-NMLZ	son CONJ daughter bear-NMLZ	letter write-NMLZ
'a cause to abandon the analysis'	'all the sons and daughters [she] bears'	'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 as quasi-nominalizers (like *-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།	(5) འགྲོ་བའི་གྲབས
dmyal-bar 'gro-nyen	*'gro-ba 'i grabs*
hell-LOC go-NMLZ	go-NMLZ GEN preparations
'danger of going to hell'	'preparations to leave'

It should be noted that all of proposition-centered nominalizers except *-Pa*, which is the most common nominalizer, also function as nouns or parts of compounds and their meanings 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, since it is often unclear whether we should interpret them as nouns (independent or parts of compounds) or as nominalizing particles. Moreover nominalized verbs in the Tibetan language are often idiomatized – acquire a specific meaning. Such verb phrases also require a certain way of modelling in the computer ontology.

1. 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]. Depending on the language, nominalizers can bear more than just a nominalizing function, they can convey an additional specific meaning (e.g., place of action) or function as converb forms [2, p. 170].

In some cases the nominalizer doesn't make the whole clause nominalized, it nominalizes only the verb. The dependents of the verb no longer require the accordance with verb transitivity and are treated as nominal dependents. This type quite often occurs in Tibeto-Burman languages and is called an «action nominalization» or «event nominalization» [2, p. 166]. In the texts of our corpus, we found cases of nominalization of a similar type, which, however, were not specifically described for the Tibetan language. Nonetheless this phenomenon causes a number of difficulties for formal grammatical and ontological modeling that will be described below.

According to S. Beyer, nominalizers in the Tibetan language can be divided 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. This type is unusual to the speakers of English, but is natural and often used in Tibetan. S. Beyer gives three patient-centered nominalizers: *-rgyu* (denotes 'patient of proposition' like in (1)), *-'o-cog/-dgu/-tshad* (denote all patients of proposition like in (2)) and *-'phro/'phros* (denote 'remainder of patient of proposition' like in (3)) [1, p. 296-298].

(1) དཔྱད་པ་གཏོང་རྒྱུ	(2) བུ་དང་བུ་མོ་བཙའོ་ཅོག	(3) ཡི་གེ་འབྲིས་འཕྲོས
dpyad-pa gtong rgyu	*bu dang bu-mo btsa'o-cog*	*yi-ge 'bri-'phros*
analyse-NMLZ abandon-NMLZ	son CONJ daughter bear-NMLZ	letter write-NMLZ
'a cause to abandon the analysis'	'all the sons and daughters [she] bears'	'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 as quasi-nominalizers (like *-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	(5) འགྲོ་བའི་གྲབས
dmyal-bar 'gro-nyen	*'gro-ba 'i grabs*
hell-LOC go-NMLZ	go-NMLZ GEN preparations
'danger of going to hell'	'preparations to leave'

It should be noted that all of proposition-centered nominalizers except *-Pa*, which is the most common nominalizer, also function as nouns or parts of compounds and their meanings

[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].

4. Real Nominalizers

4.1. Formal grammatical and Ontological Modeling of Real Nominalizers

The nominalizer *-Pa* is the most general of all nominalizers, used in a neutral context. Its form varies depending on the preceding final: *-pa* after *-g, -d, -n, -b, -m*, and *-s*; *-ba* after preceding *-r*, *-l*, and open syllables. It signals only that the entire proposition is functioning as a nominal, and contributes nothing further to the meaning of a newly formed proposition [1, p. 295].

For it the CIC NominalizerSuff was created. This class is embedded as a modifier in the classes for nominalized verbal phrases of different types (VNNoMorphon, VNNoMorphonEllArg, VNNoTenseNoMorphon, VNNoTenseNoMorphonEllArg, VNNoTenseNoMoodNoMorphon, VNNoTenseNoMoodNoMorphonEllArg).

The verbal phrase with *-Pa* can denote a process, an object or a subject of an action. Thus, one nominalized verbal phrase (e.g. 8) usually has at least two semantic versions of parsing (see fig. 1).

8) བསླབ་པ
bslab-pa
teach-NMLZ
(8.1) 'teaching'
(8.2) 'one who teaches'

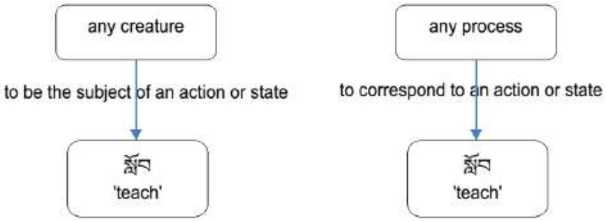
any creature
to be the subject of an action or state
སློབ
'teach'
any process
to correspond to an action or state
སློབ
'teach'

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 such nominalizers as *-rgyu*, that do not have allomorphs a common class NominalizerSuffNoMorphon was created in the formal grammar.

4.2. 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 (9), since *dang* occurs only after nouns or noun phrases [8, p. 5; 1, p. 241].

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

(10) 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 (10) 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).

The processing of prose texts will probably give enough examples to claim that zero nominalization is typical not only for poetry and can be considered as an ellipse, valid in the whole language. In this case revising of grammar will require determining 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 (11) and (12).

(11) 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 (11) 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 typical grammatical phenomena for the Tibetan language the special class PoeticHomogenVP, that was embedded into classes for verbal nominalization.

(12) 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 (12) 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.

5. Noun-nominalizers

By the term «noun-nominalizer» (i.e., quasi nominalizers) we mean nouns that can join the verb stem and act as nominalizers. In the result of the corpus data analysis the following noun-nominalizers was discovered: *mkhan* 'person', *tshul* 'way, method', *thabs* 'skill, technique', *stangs* 'manner', *rgyu* 'cause', *rtsal* 'capacity', *lugs* 'manner', *cha* 'part', *sa* 'place', *dus* 'time'. Most of them occur in three typical functions: as nouns, as elements of compounds and after verbal roots. For example, *rgyu* can perform all these functions and there are no ambiguous cases. However most of the other potential nominalizers can be found in contexts that do not clearly define their status.

5.1. Nominalizer *mkhan* 'a person, skillful in'

In the corpus *mkhan* (variant of the noun morpheme *mkhan po*) is used mainly as nominalizing particle, as shown in examples (13) and (14).

(13) འབྲི་མཁན་རྣམས
'bri mkhan rnams
write-NMLZ-PL
'[those, who are] skillful in writing'

(14) གདམས་ངག་སྟོན་མཁན་གྱི་སློབ་དཔོན་དེ
gdams-ngag ston-mkhan-gyi slob-dpon de
instruction teach-NMLZ-GEN teacher DEM
'this teacher, [who is] skillful in teaching instructions'

In some cases *mkhan* occurs after a noun, as in the example (16). Such cases we model as compounds.

(15) ཡི་གེའི་མཁན་པོ
yi-ge 'i mkhan-po
letter GEN sage
'a person well-versed in letters'

(16) ཡིག་མཁན
yig-mkhan
letter_sage
'a literate person'

All Tibetan compounds are created by the juxtaposition of two existing words. Compounds are virtually idiomatized contractions of syntactic groups which have inner syntactic relations frozen and are often characterized by omission of grammatical morphemes [1, p. 102]. In the example (16) *mkhan* is the component of the genitive compound (the compound version for the full word *mkhan po* 'sage'). This class of compounds is NPGenCompound, a frequent type of Tibetan compounds derived from the noun phrases with genitive arguments by omission of the genitive case marker [6, p, 149].

The relation between NPGenCompound components is subordinate genitive relation. When modeling compounds of this type in the computer ontology, it is necessary not only to model meaning of the compound and both compound components, but also to establish specific subclass of the general genitive relation 'to have any object or process (about any object or process)' between basic classes of compound components. For example, NPGenCompound (16) was formed from the genitive nominal group (15), consisting of two nouns. Thus, the head component of the compound is NRoot *yig* 'letter'; and the argument is NPGenCompoundArg that consists of the head immediate constituent (the second NRoot *mkhan* 'sage'), attached with the intersyllabic delimiter (argument immediate constituent) on the left. To ensure the correct semantic parsing of the compounds of this type the concept 'any object' (the basic class for the first component of the compound (16) – *yi-ge* 'letter') had to be connected with the concept *mi* 'person', which is a basic class of the second component, with a relation 'to have a person well-versed in something (about any object)', which is a subclass of the general genitive relation.

5.2. Nominalizer *tshul* 'way'

It is a commonly used noun that can occur as independent noun after verbs that are already nominalized like in (17) (in this case the verb is nominalized with -*Pa* and the whole phrase is put in the genitive case) or as a part of compounds of different types (for e.g. (18)).

(17) དོན་ལ་དབབ་པའི་ཚུལ	(18) ཚུལ་ཁྲིམས
don la dbab-pa 'i tshul	*tshul khrims*
meaning LOC descend-NMLZ GEN way	way_law
'way of gaining an insight into the meaning'	'mode of conduct'

Its status as a standard clausal nominalizer cannot be considered unambiguous, since *tshul* can be used directly after verbs in different ways in the same text and even almost in the same phrase. For example, in (19) one can see the standard case of clausal nominalization–the transitive verbal phrase *yi-ge 'bri* 'write letters' is nominalized with the use of -*tshul.*

(19) ཡི་གེ་འབྲི་ཚུལ་སྦྱང	(20) ཡི་གེའི་འབྲི་ཚུལ་བཤད
yi-ge 'bri tshul sbyang	*yi-ge 'i 'bri tshul bshad*
letter write-NMLZ train	letter GEN write_way explain
'train in the way of writing letters'	'explain the way of writing of letters'

However, in (20) the verb loses its transitivity as the direct object in this verbal phrase is used in the genitive case that is typical for action nominalization.

Such cases cause difficulties in semantic modeling. If we consider (20) a case of nominalization than we should “allow” letters to have processes in the computer ontology, that is to connect the concept '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. We also can interpret the verb *'bri* 'write' and *tshul* here as noun phrase with genitive compound which with its right context form a genitive noun phrase. In this case the head component of the compound is CIC CompoundAtomicVN *'bri* 'write' that stands for compound atomic nominalized verb within a compound (the nominalizer in compounds is always omitted, thus, the nominalized verb form superficially comprises the verb root only), while *tshul* here is treated as noun – the head class of the NPGenCompoundArg. In the computer ontology the subclass of the general genitive relation 'to have a way (about any process)' should be set between the class 'any process' and *tshul* 'way' which is the basic class itself. Unfortunately, both approaches require the same effort from an ontology editor and potentially cause semantic ambiguity.

5.3. Nominalizer *thabs* 'skil, technique'

Same as *tshul*, *thabs* can be used after nominalized verbs (21) or both as a standard clausal nominalizer (22) and as an action nominalizer (23).

(21) དམག་བྱེད་པའི་ཐབས	(22) སྔགས་ཡིག་ཀློག་ཐབས	(23) སྣག་ཤོག་སྨྱུག་གསུམ་གྱི་བཟོ་ཐབས
dmag byed-pa 'i thabs	*sngags-yig klog thabs*	*snag shog smyug gsum-gyi bzo-thabs*
war make-NMLZ GEN skill	mantra read-NMLZ	ink paper pen three GEN make-skill
'a skill of making war'	'a skill of reading mantra'	'a skill of making three [items]: ink, paper and pen'

In the examples (22) and (23) *thabs* is used after the verbs *klog* 'read' and *bzo* 'make'. Both of them have the same valency – the subject in the ergative case and the direct object in the absolutive case. In the example (22) the direct object is *sngags-yig* 'mantra'. It is used in the absolutive case, that allows to interpret *thabs* as the nominalizer for the verbal phrase *sngags-yig klog* 'to read a mantra'. However, in the example (23) the initial verbal phrase is not preserved, since the direct objects of the verb *bzo* 'make' (*snag shog smyug gsum* 'ink, paper and pen') are used with the genitive case marker.

5.4. Nominalizer *lugs* 'method'

The noun *lugs* can be used as a noun as in example (24) in the genitive noun phrase with the verb phrase nominalized with *-Pa*. Occasionally it can be used as a clausal nominalizer as in (25), where the valency of the verb *mthun* 'comply' that requires an object in the associative case is preserved. It also can derive a noun from a verb which functions as a head of a noun phrase (26).

(24) ཇི་ལྟར་མཐུན་པར་སྦྱོར་བའི་ལུགས་	(25) ཨི་དང་མཐུན་ལུགས	(26) ཡི་གེའི་སྐུང་ལུགས
ji-ltar mthun-par sbyor pa'i lugs	*i dang mthun-lugs*	*yi-ge'i skung lugs*
how correspond-NMLZ LOC	i ASS	letter GEN to
add-NMLZ GEN method	correspond-NMLZ	shorten_method
'way of corresponding adding'	'letter *i* matching rules'	'the method of shortened [writing] of letters'

5.5. Nominalizer *sa* 'place'

It is the last noun-nominalizer, indicated by S. Beyer that was found in the corpus. Unlike previous nouns *sa* is not used as an action nominalizer in the corpus texts. Still there are some cases when there is no left context or the context can’t help to define whether it is a nominalizer or a part of compound.

Nouns *stangs* 'manner', *rtsal* 'capacity', *cha* 'part' and *dus* 'time' were not mentioned by S. Beyer as nominalizers. However they occur in the same contexts. For example, the noun *stangs* 'manner' in the example (27) functions as clausal nominalizer and the verb valency is preserved, but in the example (28) the verb *'bri* 'write' and *stangs* are preceded by a noun in the genitive case (that is the direct object of the verb).

(27) ལ་དོན་གྱི་ཕྲད་སྦྱར་སྟངས	(28) མིང་གི་འབྲི་སྟངས
la-don gyi phrad sbyar stangs	*ming gi 'bri stangs*
la-meaning GEN particle add-NMLZ	name GEN write manner
'rules of the use of particles with the meaning of la'	'manner of writing the name'

Since all nouns discussed in this section 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. The CIC NominalizerNRoot was added into the formal grammar with noun root 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 for all nouns that occurs as nominalizers in our corpus the same basic class was created. This class was connected with a specific relation with the basic class for verbs so that meanings of nouns were reflected in semantic versions of parsing.

6. Idiomatization 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 (29) and (30).

(29) བྱེད་པ	(30) མཁས་པ
byed-pa	*mkhas-pa*
do-NMLZ	be_learned-NMLZ

'semantic role'	'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 concepts with the meanings 'doer' and 'doing' are created for the expression (29), and concepts with the meanings 'one who knows' and 'knowing' are created for the expression (30).

Verbs and verbal phrases formed by noun nominalizers (or quasi-nominalizers) also can be idiomatized. Such cases were also discovered in the corpus. For example, (31) and (32) are grammatical terms formed by nominalization of two transitive verbal phrases. If an idiom is represented by a single verb and noun nominalizers (e.g., (33)) it is modeled as a compound.

(31) ལ་དོན་སྦྱོར་ཚུལ	(32) བྱེད་སྒྲ་སྦྱོར་ཚུལ	(33) ཀློག་ཚུལ
la-don sbyor-tshul	*byed-sgra sbyor-tshul*	*klog-tshul*
la_meaning add-NMLZ	do_marker add-NMLZ	read_way
'rules of the use of particles with the meaning of *la*'	'rules of the use of agent marker'	'transcription'

Thus, *klog-tshul* (33) 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.

7. 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	17

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.

Conclusions and Further Work

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.

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».

References

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

[2] Genetti C. Nominalization in Tibeto-Burman languages of the Himalayan area: A typological perspective // Nominalization in Asian Languages: Diachronic and typological perspectives, edited by Foong Ha Yap, Karen Grunow-Hårsta and Janick Wrona. 2011. P. 163–194.

[3] Dobrov A., et al. Morphosyntactic analyzer for the Tibetan language: aspects of structural ambiguity / Dobrov A., Dobrova A., Grokhovskiy P., Soms N., Zakharov V. // International Conference on Text, Speech, and Dialogue. 2016. P. 215-222. https://doi.org /10.1007/978-3-319-45510- 5_25.

[4] Dobrov A., et al. Morphosyntactic Parser and Textual Corpora: Processing Uncommon Phenomena of Tibetan Language / Dobrov A., Dobrova A., Grokhovskiy P., Soms N. // Proceedings of the International Conference IMS. 2017. P. 143–153.

[5] Dobrov A., et al. Computer Ontology of Tibetan for Morphosyntactic Disambiguation / Dobrov A., Dobrova A., Grokhovskiy P., Smirnova M., Soms N. // Digital Transformation and Global Society. DTGS. Communications in Computer and Information Science, Vol. 859, edited by Alexandrov D., Boukhanovsky A., Chugunov A., Kabanov Y., Koltsova O. Springer, Cham, 2018. P. 336–349. https://doi.org/10.1007/978-3-030-02846-6_27.

[6] Dobrov A., et al. Idioms Modeling in a Computer Ontology as a Morphosyntactic Disambiguation Strategy / Dobrov A., Dobrova A., Grokhovskiy P., Smirnova M., Soms N. // Text, Speech, and Dialogue. TSD 2018. Lecture Notes in Computer Science, Vol. 11107, edited by Sojka P., Horák A., Kopeček I., Pala K. Springer, Cham, 2018. P. 76–83. https://doi.org/10.1007/978-3-030- 00794-2_8.

[7] Dobrov A., et al. Formal grammatical and ontological modeling of corpus data on Tibetan compounds / Dobrov A., Dobrova A., Smirnova M., Soms N. // IC3K 2019 - Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. Vol. 2. SciTePress, 2019. P. 144–153.

[8] Hill N.W. Tibetan zero nominalization // Revue d'Etudes Tibétaines. 2019. № 48. P. 5–9.

[9] AIIRE Ontology. URL: http://svn.aiire.org/repos/ontology/ (access date: 10.06.2020).

[10] AIIRE Ontohelper. URL: http://svn.aiire.org/repos/ontohelper/ (access date: 10.06.2020).

[11] Tibetan ontology. URL: http://svn.aiire.org/repos/tibet/trunk/aiire/lang/ontology/concepts.xml (access date: 10.06.2020).

[12] Tibetan ontology (unauthorized view). URL: http://ontotibet.aiire.org (access date: 10.06.2020).