Formal Grammatical and Ontological Modeling of Corpus Data on Tibetan Compounds

Aleksei Dobrov[1][a], Anastasia Dobrova[2][b], Maria Smirnova[1][c] and Nikolay Soms[2][d]
[1]*Saint-Petersburg State University, Saint-Petersburg, Russia*
[2]*LLC "AIIRE", Saint-Petersburg, Russia*

Keywords: Tibetan Language, Compounds, Computer Ontology, Tibetan Corpus, Natural Language Processing, Corpus Linguistics, Immediate Constituents.

Abstract: This article provides a consistent formal grammatical and ontological description of the model of the Tibetan compounds system, developed and used for automatic syntactic and semantic analysis of Tibetan texts, on the material of a hand-verified corpus. This model covers all types of Tibetan compounds, which were previously introduced by other authors, and introduces a number of new classes of compounds, taking into account their derivation, structure and semantics. The article describes the tools used for ontological modeling of Tibetan compounds; special attention is paid to the problem of modeling the semantics of verbs and verbal compounds. Nominal and verbal compounds are considered separately, it is noted that the importance of verbal compounds for the Tibetan language system is not less than that of nominal compounds. The statistical data on the absolute frequency distribution of the use of compounds of different types in the current version of the corpus annotation and on the amounts of ontology concepts associated with each class of compounds are given.

1 INTRODUCTION

The research introduced by this paper is a continuation of several research projects ("The Basic corpus of the Tibetan Classical Language with Russian translation and lexical database", "The Corpus of Indigenous Tibetan Grammar Treatises", "Semantic interpreter of texts in the Tibetan language"), aimed at the development of a full-scale natural language processing and understanding engine based on a consistent formal model of Tibetan vocabulary, grammar, and semantics, verified by and developed on the basis of a representative and hand-tested corpus of texts.

The Basic Corpus of the Tibetan Classical Language (The Basic Corpus of the Tibetan Classical Language, 2019) and the Corpus of Indigenous Tibetan Grammar Treatises (The Corpus of Indigenous Tibetan Grammar Treatises, 2019) comprise 34,000 and 48,000 tokens, respectively. Tibetan texts are represented both in the Tibetan Unicode script and in a standard Latin (Wylie) transliteration (Grokhovskii et al., 2015). These corpora are developed, annotated and tested manually by a team of professional tibetologists, and in this sense are unique today.

The ultimate goal of the current project is to create a formal model (a grammar and a linguistic ontology) of the Tibetan language, including morphosyntax, syntax of phrases and hyperphrase unities, and semantics that can produce a correct morpho-syntactic, syntactic, and semantic annotation of the corpora without any manual corrections.

This article discusses the results achieved currently in modeling Tibetan compounds, both from syntactical and from semantical perspective.

In Tibetan, there is no clear boundary between morphology and syntax; at least, there are no materially expressed boundaries of word forms and, from the point of view of an automatic system, the analysis of compounds is in no way different from the analysis of free combinations of Tibetan morphemes like noun phrases or sentences. Modeling compounds is one of the most important tasks in the current research, not only because the frequency of use of compounds in Tibetan texts is high (at least, as compared with texts in Indo-European languages), but also because without a correct syntactic and semantic model of compounds, a huge ambiguity of Tibetan text segmentation and parsing arises, which leads to a combinatorial

[a] https://orcid.org/0000-0003-0245-5407
[b] https://orcid.org/0000-0002-8419-1005
[c] https://orcid.org/0000-0001-5429-2051
[d] https://orcid.org/0000-0002-0546-5101

Dobrov, A., Dobrova, A., Smirnova, M. and Soms, N.
Formal Grammatical and Ontological Modeling of Corpus Data on Tibetan Compounds.
DOI: 10.5220/0008162401440153
In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019)*, pages 144-153
ISBN: 978-989-758-382-7

explosion (Dobrov et al., 2018a, p. 345).

To date, there is no systematic description, even no single classification of Tibetan compounds in tibetological literature, which could be considered generally accepted. The classification, which is given in this article for the first time, was developed on the basis of formal-grammatical and ontological modeling of the phenomena observed in the above-mentioned corpora, and does not pretend to be universal, however, it covers all types of Tibetan compounds, which were introduced by the researchers earlier, and introduces new classes which do not seem to be earlier described.

The presented classification also covers all types of compounds of the corpora. The exceptions are few specific cases found in the poetic tests. However, since Tibetan poetic texts are characterized by a large number of omissions of grammatical markers and specific abbreviations, the status of these lexical units remains to be determined. It should be also noted that the corpora include only texts in the classical Tibetan language. Modern Tibetan texts are planned to be added to the corpus, which may lead to the discovery of new types of compounds. In this case the current version of grammar will be corrected.

Unlike many other languages, the Tibetan language is characterized by wide use not only of nominal, but also of verbal compounds derivational models. Modeling verbal compounds is more complicated for some further-mentioned reasons than modeling nominal compounds; therefore, this article considers not only the results of this modeling, but also the tools that were used to obtain them.

2 RELATED WORK

Linguistic ontologies in natural language understanding (NLU) systems are used as analogues for the semantic dictionaries that were used before (cf. (Melcuk, 1995); (Mel'čuk and Žolkovskij, 1984), (Leont'eva, 2003) etc.); the main difference between an ontology and a conceptual dictionary is that, in a semantic dictionary, semantic valencies are, in fact, postulated, whereas in ontologies, valencies are automatically computed by inference engine subsystems; semantic restrictions are defined not in terms of word lists, but in terms of base classes of ontological concepts (that is the idea behind the mechanism of word-sense disambiguation in (Dobrov et al., 2016), (Dobrov, 2014), (Matuszek et al., 2006), (Rubashkin et al., 2012), etc.).

As a formal model, ontology has to predict permissible and exclude impermissible relations between concepts. Despite the clearness and obviousness of these two requirements, there is no generally accepted definition of the term 'ontology' in the scientific literature, which would have reflected them. The most famous and widely cited general definition of the term 'ontology' is 'an explicit specification of a conceptualization' by Gruber (Gruber, 1993). Many different attempts were made to refine it for particular purposes. Without claiming for any changes to this defacto standard, we have to clarify that, as the majority of researchers in natural language understanding, we mean not just any 'specification of a conceptualization' by this term, but rather a computer ontology, which we define as a database that consists of concepts and relations between them.

Ontological concepts have attributes. Attributes and relations are interconnected: participation of a concept in a relation may be interpreted as its attribute, and vice versa. Relations between concepts are binary and directed.

They can be represented as logical formulae, defined in terms of a calculus, which provides the rules of inference. Relations themselves can be modeled by concepts.

There is a special type of ontologies - so called linguistic ontologies ((Dobrov et al., 2016), (Dobrov, 2014), (Matuszek et al., 2006), (Rubashkin et al., 2012), etc.), which are designed for automatic processing of unstructured natural language texts. Units of linguistic ontologies represent concepts behind meanings of real natural language expressions. Ontologies of this kind actually model linguistic picture of the world that stands for language semantics, as subject domain. Ontologies that are designed for natural language processing are supposed to include relations that allow to perform semantic analysis of texts and to perform lexical and syntactic disambiguation. The ontology, used for this research, was developed according with the above mentioned principles (Dobrov, 2014). Its structure is described in detail in the articles (Dobrov et al., 2018a), (Dobrov et al., 2018b). Totally within the framework of this research 4335 concepts that are meanings of 3943 Tibetan expressions were modelled in the ontology.

The only attempt to classify Tibetan compounds was made by Stephan V. Beyer. 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 (Beyer, 1992, p. 102). E.g., phrase (1) is clipped to (2).

Depending on part of speech of compound and its components Stephan V. Beyer identifies several mod-

(1) ཁའི་རྒྱན་	(2) ཁ་རྒྱན་
kha ’i rgyan	kha rgyan
mouth GEN adornment	mouth_adornment
’adornment of mouth’	’moustache’

els of compound-building in the Tibetan language. The following five models are original Tibetan: noun + noun → noun; noun + adjective → noun; adjective + noun → noun; adjective + adjective → noun; noun + verb → verb (Beyer, 1992, p. 103–105). Stephan V. Beyer also notes that the Tibetan language uses additional devices for compound-building, in part borrowed by Tibetans from Sanskrit within the process of translation: noun + verb → noun; intensifier + verb → verb (Beyer, 1992, p. 108–110).

According to syntactic relation between the components Tibetan compounds may be divided into two main classes: compounds with subordinate relations and compounds with coordinate relations (Grokhovskii and Smirnova, 2017, p. 137).

3 THE SOFTWARE TOOLS FOR PARSING AND FORMAL GRAMMAR MODELING

This study was performed with use of and within the framework of the AIIRE project (Dobrov et al., 2016). 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 (Dobrov et al., 2017) 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:

- The template of semantic graph which represents the meaning of this constituent;
- The list of possible head constituent classes;
- The list of possible subordinate constituent classes;
- The dictionary of possible linear orders of the subordinate constituent in relation to the head and the meanings of each order;
- The boolean field for head ellipsis possibility;
- The boolean field for subordinate constituent ellipsis possibility;
- 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.

4 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 (http://svn.aiire.org/repos/ontology/, http://svn.aiire.org/repos/ontohelper/), and the ontology itself is available as a snapshot at http://svn.aiire.org/repos/tibet/trunk/aiire/lang/ontology/concepts.xml and it is also available for unathorized view or even for edit at http://ontotibet.aiire.org (edit permissions can be obtained by access request). The basic ontological editor is described with examples from the Tibetan ontology in (Dobrov et al., 2018a), (Dobrov et al., 2018b),

(Grokhovskii and Smirnova, 2017).

4.1 Ontological Editor for Modeling Verb Concepts in the Ontology

Modeling verb (or verbal compound) 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 are used to speed up and partly automate verbal concepts modeling. AIIRE ontological editor - Ontohelper is used to build the whole hierarchy of superclasses for any verb meaning in the ontology.

The logic behind this tool is also based on the division of verbs into dynamic (terminative and non-terminative) and static ones (Maslov, 1998). Dynamic verbs express actions, events and processes associated with different changes. Static verbs express states, relations or qualities (GED, , p. 105). A terminative verb denotes an action which has a limit in its development. A non-terminative verb denotes an action which doesn't admit of any limit in its development (activity).

When using the Ontohelper editor, it is necessary to determine whether the verb being modeled denotes action, state or activity. Terminative, non-terminative and static verb meanings correspond to subclasses of concepts 'to perform an action', 'to perform an activity' and 'to be in a state' in the ontology, respectively. The basic class for subjects of the verb to be modeled is indicated, as well as the basic class of direct objects for transitive verbs and the class of indirect dative objects for verbs denoting addressed actions. It is also possible to specify classes of circumstances, i.e., objects with special case government (e.g., for verbs that govern the associative case).

When all necessary attributes of a verb meaning are specified, the Ontohelper editor builds the whole ontological classes hierarchy from scratch for this particular combination of attributes, and if some classes are already present in the ontology, they are not built again, but tested in terms of consistency with the current actant / circumstant relations model. This allows not only to boost the speed of semantic valencies fine-tuning for verb classes, but also to rebuild the whole hierarchy in cases when new actant / circumstant relation or class has to be established according to some new observations on the corpus phenomena.

5 CLASSIFICATION OF TIBETAN COMPOUNDS AND THEIR MODELING IN THE FORMAL GRAMMAR AND THE COMPUTER ONTOLOGY

Depending on the part of speech classification, nominal and verbal compounds can be distinguished. Initially, the ontology allowed binding concepts to expressions with marking the expression as an idiom and establishing a separate type of token, common for nominal compounds. Since a large number of combinatorial explosions were caused by the incorrect versions of compounds parsing (the same sequence of morphemes can be parsed as compounds of different types) and their interpretation as noun phrases of different types, it was decided to expand the number of token types in the ontology according to identified types of nominal and verbal compounds (see below).

As all Tibetan compounds are idioms, in AIIRE ontology, in addition to the meanings of a compound, meanings of its components are also modeled, so that they could be interpreted in their literal meanings too. This is necessary, because AIIRE natural language processor is designed to perform natural language understanding according with the compositionality principle (Pelletier, 1994), and idiomaticity is treated not merely as a property of a linguistic unit, but rather as a property of its meaning, namely, as a conventional substitution of a complex (literal) meaning with a single holistic (idiomatic) concept (Dobrov et al., 2018b, p. 78–79).

5.1 Nominal Compounds Modeling in the Formal Grammar and the Computer Ontology

Depending on the syntactic model of the compound derivation, the following types were distinguished for nominal compounds: compound noun root group (CompoundNRootGroup); compound attribute group (CompoundAttrGroup); noun phrase

with genitive compound (NPGenCompound); compound class noun phrase (CompoundClassNP); adjunct compound (AdjunctCompound); named entity compound (NamedEntityCompound).

Compound noun root group (3) consists of NRoot (nominal root), being the head class, and CompoundNRootGroupArg, attached as a subordinate constituent. The linear order of the subordinate constituent in relation to the head is right. CompoundNRootGroupArg stands for compound argument that consists of a noun root attached with an intersyllabic delimiter - upper dots (Tib. tshegs), like in (3). NForeign (foreign noun) is allowed to be head in both CICs - CompoundNRootGroup and CompoundNRootGroupArg. The linear order of the CompoundNRootGroupArg subordinate constituent in relation to the head is left. Heads and arguments of all Tibetan compounds can not be ellipsed. For all compounds the setting 'only_idiom=True' was also made. According to this setting any non-idiomatic interpretations of a compound are excluded.

This type of compounds does not require establishment of any semantic relations in the computer ontology for each compound. It is enough that the meaning of the compound and its components are modeled in the ontology, and that the general coordination mechanism is also modeled in the module for syntactic semantics (the meaning of a coordinate phrase is calculated as an instance of 'group' concept which involves 'include' relations to its elements). Compound attribute groups (4) also belong to this semantic type. It is a group of superficially homogeneous attributes within a compound. This way of derivation is quite frequent for Tibetan personal names (the name consists of a set of epithets (attributes), without any explicit noun) and for words, denoting size (e.g. (4)). CompositeAttrGroup consists of CompositeAtomicAttributeTopic and CompositeAttrCoord, where the first part is the first attribute or group of attributes and the second part is the last attribute attached as a subordinate constituent. If there are more than two attributes, they are attached in exactly the same way with CompositeAttrGroup self-embedding. The morphosyntactic structure of compounds of this type is described in detail in (Dobrov et al., 2017).

(3) གཙུག་ལག་	(4) རིང་ཐུང་
gtsug-lag	ring-thung
crown_of_head_hand	long_short
'basket'	'length'

NPGenCompound is another frequent class of Tibetan nominal compounds. These compounds are derived from noun phrases with genitive arguments by omission of the genitive case marker. In accordance with the current grammar version, the head constituents of NPGenCompound can be CompoundAtomicNP, NForeign, PlaceNameForeign, LetterCnt (countable letter), PDefRoot, PIntRoot, NPGenCompound; the subordinate constituent class can only be NPGenCompoundArg. The linear order of the subordinate constituent in relation to the head is right.

CompoundAtomicNP means atomic nominal phrase within a compound. PlaceNameForeign (foreign place name) was allowed to be the head in NPGenCompound for such cases as e.g. (5). PDefRoot (definitive pronoun) and PIntRoot (indefinite pronoun) were included into possible head classes for such combinations as e.g. (6).

(5) སི་ཁྲོན་མི་རིགས་	(6) གཞན་དབང་
si-khron-mi-rigs	gzhan-dbang
Sichuan-nationality	other-power
'Sichuan people'	'dependent connector'

In some cases one of the components of a compound is itself a compound. For example, in the NPGenCompound (7) the head class can also be NPGenCompound (8). This class of immediate constituents is not the only case when a compound is among heads or arguments of another compound. In the poetic texts, even more complex structures were discovered, the status of which is still to be clarified.

(7) དཔེ་མཛོད་ཁང་	(8) དཔེ་མཛོད་
dpe-mdzod-khang	dpe-mdzod
book-repository-house	book-store
'library'	'book repository'

The CIC NPGenCompoundArg stands for a genitive compound argument that consists of the head immediate constituent, attached with the intersyllabic delimiter (argument immediate constituent) on the left. Head classes of NPGenCompoundArg include: CompoundAtomicVNNoTenseNoMood, IndepNRoot, OnlyCompoundNRoot, NForeign, PersNameForeign (personal name foreign), NPGenCompound, CompoundAtomicVN. CompoundAtomicVN stands for an atomic nominalized verb within a compound (the nominalizer in compounds is always omitted, thus, the nominalized verb form superficially comprises the verb root only). CompoundAtomicVNoTenseNoMood is a CompoundAtomicVN which does not have neither mood, nor tense. As in this case, Tibetan verb roots often do not have different allomorphs for different moods and tenses. IndepNRoot (independent noun root) is a noun root (allomorph of a noun root), which

can be used both within a compound and in free combinations. OnlyCompoundNRoot stands for a noun root (allomorph of a noun root), which can be used only within a compound (such noun roots are often single-syllabic contractions of multisyllabic roots).

The relation between NPGenCompound components is subordinate genitive relation. When modeling compounds of this type in the computer ontology, it is necessary to establish specific subclasses 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 (9) was formed from the genitive nominal group (10). Thus, the concept 'geographical object' (the basic class for the first component of the compound (9) - bod 'Tibet') had to be connected with the concept skad 'language', which is a basic class itself, with a relation 'to have a language (about any geographical object)', which is a subclass of the general genitive relation.

(9) བོད་སྐད་	(10) བོད་ཀྱི་སྐད་
bod-skad	bod gyi skad
Tibet_language	Tibet GEN language
'the Tibetan language'	'the language of Tibet'

Another frequent class of Tibetan nominal compounds is CompoundClassNP. Compounds of this type are derived from regular noun phrases with adjectival or, more often, quasi-participial (there are no participles in the Tibetan language, but rather nominalized verbs that can act both as participles and as processual nouns) attributes. Possible head classes for the CIC CompoundClassNP are IndepNRoot, OnlyCompoundNRoot, NForeign, PersNameForeign (foreign personal name), PlaceNameForeign, LetterCnt. Itsmodifier class can be CompoundAtomicAttribute, CompoundAtomicAttributeNoTense, CompoundAtomicAttributeNoTenseNoMood. CompoundAtomicAttribute consists of a state verb, denoting an object feature (the head class), attached by the intersyllabic delimiter (argument class) on the left. For example, CompoundClassNP (11) has the head class IndepNRoot srog 'breath' and the modifier class CompoundAtomicAttributeNoTenseNoMood, consisting of the intersyllabic delimiter attached on the left to the verb chen 'be big' which does not have neither mood, nor tense.

The only requirement for modeling compounds of this type in the ontology is that the basic class of a nominal component in a compound must be a subclass of the specified verb subject (for the verbal component of the compound). I.e., the verb, from which the attribute is derived, must allow this subject by its valencies.

Another class of Tibetan nominal compounds is Adjunct compound (e.g. (12)). Compounds of this class are derived from regular noun phrases with adjuncts. Thus, the head class for this CIC is NRoot; and the argument class is CompoundRightNRootArg, consisting of a noun root and the intersyllabic delimiter. It is necessary that the components of the compound belong to the same basic class in the ontology, or that there is no limitation on their equivalence relations (the classes of the concepts should not be disjoint).

(11) སྲོག་ཆེན་	(12) ལེ་ཚན་
srog-chen	le-tshan
breath_be_big	section_section
'aspiration'	'section'

Three upper-mentioned classes of nominal compounds have exactly the same surface structures: NRootGroupCompound, NPGenCompound, and AndjuctCompound. Compounds of all three classes look like combinations of two bare noun roots, but they have completely different syntactic structures and completely different semantic models. As all compounds are modeled as idioms, when binding ontology concepts, that they denote, to Tibetan language units, it is necessary to specify the syntactic class (type of token) for each concept and to make the natural language processing engine exclude all other possible parses thereof.

Finally, there is also such class of nominal compounds as NamedEntityCompund. This class was introduced for combinations of letters or exponents of arbitrary Tibetan morphemes with NRoot like in (13). It was decided that the Letter or Exponent is the head component of NamedEntityCompound.

(13) ལ་སྒྲ་
la-sgra
la_marker
'grammatical marker la'

The NamedEntityCompound CIC is a subclass of named-entity nomination, where the name of the entity is a letter or an exponent of any Tibetan morpheme. Thus, semantic restrictions are imposed on the possible classes of the subordinate constituent concepts, due to the fact that only linguistic units can have such names according with the ontology (there is a corresponding relation between the 'linguistic unit' and 'linguistic unit exponent' concepts).

5.2 Verbal Compounds Modeling in the Formal Grammar and the Computer Ontology

Depending on the syntactic model of the compound derivation, the following types were distinguished for verbal compounds: verb coordinate compound (VerbCoordCompound); compound transitive verb phrase (CompoundTransitiveVP); compound atomic verbal phrase with circumstance (CompoundAtomicVPWithCirc) and compound associative verb phrase (CompoundAssociativeVP).

In fact, each of these types of verbal compounds is represented by three types in the current grammar version - verbal compound, which varies tense and mood (e.g. CompoundTransitiveVP); verbal compound, which varies only in mood (e.g. CompoundTransitiveVPNoTense); and verbal compound, which doesn't vary in tense and mood (e.g. CompoundTransitiveVPNoTenseNoMood). Verbal compounds like other verbs are processed using the Ontohelper editor.

VerbCoordCompound (e.g. (14)) consists of VRoot (verbal root, being the head of VerbCoordCompound) and VerbCompoundCoord that stands for the second verb (VRoot being the head of VerbCompoundCoord) with the intersyllabic delimiter. Modeling a verb coordinate compound meaning does not require establishing any special semantic relations in the computer ontology, because the upper-mentioned general coordination meaning evaluation is involved. These compounds are contractions of regular coordinate verb phrases with conjunctions omitted.

In compound transitive verb phrase (15), the first nominal component is a direct object of the second verbal component. The head class thisCompoundTransitiveVP can be VRoot or CompoundTransformativeVP. The arguments include CompoundInstanceNPArg and CompoundAtomicVNArg. The linear order of the subordinate constituent in relation to the head is left.

CompoundTransformativeVP is a contraction of a regular Tibetan transformative verb phrase, i.e., a verb phrase with terminative object. As the corpus shows, compound transformative verb phrases can themselves be parts of compound transitive verb phrases, i.e., the complete compound can be a contraction of a verb phrase both with terminative and absolutive objects.

Heads of CompoundInstanceNPArg, that is a noun phrase argument within a compound, are IndepNRoot, OnlyCompoundNRoot, PIndRoot and PIntRoot. The argument class is represented by the intersyllabic delimiter.

CompoundAtomicVNArg stands for a compound argument that consists of CompoundAtomicVN, CompoundAtomicVNNoTense, CompoundAtomicVNNoTenseNoMood being head classes, and intersyllabic delimiter argument.

To ensure the correct analysis of compounds of this type, it is necessary that the concept of the nominal component of the compound be a subclass of the basic class specified as a direct object class for the concept of the verbal component of the compound. E.g., the literal meaning of the compound (15) is 'to fasten help'. The class 'any object or process', which includes the concept phan-pa 'help', was specified as a direct object for the verb 'dogs 'to fasten'.

(14) སངས་རྒྱས་	(15) ཕན་འདོགས་
sangs-rgyas	phan-'dogs
be_purified_be_broaden	help_fasten
'awaken and broaden'	'assist'

The CIC CompoundAtomicVPWithCirc was made for a combination of CompoundAtomicVP (verbal phrase within a compound represented by a single verb root morpheme – the head class) and the modifier – CompoundCircumstance, attached on the left. CompoundCircumstance stands for a terminative noun phrase within a compound, consisting of on atom (CompoundAtomicTerminativeNP) and the intersyllabic delimiter (the terminative case marker is omitted as usually in compounds).

The basic class of the nominal component of CompoundAtomicVPWithCirc should be connected by the relation 'to be a relationship object' with the relation 'to have a manner of action or state' - the terminative case meaning. Thus, for the compound (16), this relation was established on the basic class of its nominal component rnam-pa 'type' – 'any category'.

The texts, as a rule, use the idiomatized nominalized forms of verbal compounds with the omission of the syllabic formative –pa (a nominalizer). Thus, the nominalized form of the verbal compound rnam-dbye denotes a grammatical term 'case'. In this regard, in addition to the verbal compound (16), its full nominal form (17) is also processed in the computer ontology.

(16) རྣམ་དབྱེ་	(17) རྣམ་པར་དབྱེ་བ་
rnam-dbye	rnam-pa r dbye-ba
type_divide	type LOC divide-NMLZ
'divide into classes'	'case'

CompoundAssociativeVP is another class of Tibetan verbal compounds which was introduced for contractions of regular associative verb phrases. It consists of the associative verb (the head class) and

its indirect object (possible arguments being CompoundInstanceNPArg, CompoundAtomicVNArg).

Thus, the first component of the compound (18) lhag-ma 'remainder' should belong to the class of associative objects specified for the verb bcas 'to possess' in the Ontohelper editor. Full idiomatized nominal form of this compound (19) is also modeled in the computer ontology.

(18) ལྷག་བཅས་	(19) ལྷག་མ་དང་བཅས་པ་
lhag-bcas	lhag-ma dang bcas-pa
remainder_possess	remainder ASS possess-NMLZ
'have a continuation'	'continuative'

In most cases, the direct hypernym of verbal compounds is the concept expressed by their verbal component. For example, verbal compounds (20) and (21) have the same hypernym, that is their verbal component 'chad 'explain'.

(20) གོང་བཤད་	(21) རྣམ་བཤད་
gong-bshad	rnam-bshad
top_explain	type_explain
'explain above'	'explain completely'

In other cases, there is no class-superclass relation between the meaning of the verbal compound and the verb from which it is derived. However, their type and valency are always the same.

Moreover, it was revealed that such grammatical features of Tibetan verb compounds as transitivity, transformativity, dativity, and associativity are always inherited from the main verb, even when the corresponding syntactic valency seems to be fulfilled within the compound.

6 CURRENT CORPUS ANNOTATION AND ONTOLOGICAL STATISTICS

Absolute frequencies of compounds use in current corpora annotation are represented in Table 1. This data shows that the most frequent Tibetan compound class is NPGenCompound, which is a nominal compound, but already the second most frequent compound class is CompoundAtomicVPWithCirc, which is verbal. That means that verbal compounds are not less important for the Tibetan language system than nominal ones.

The current amounts of ontological concepts for each compound class are represented in Table 2 for nominal compounds and in Table 3 for verbal compounds. This data shows the similar distribution among compound classes and, in some respect, reflects the productivity of the compound derivational models.

Table 1: Statistics on compound use in current corpora.

Compound class	Absolute frequency
NPGenCompound	1581
CompoundAtomicVPWithCirc	718
CompoundNRootGroup	465
AdjunctCompound	333
CompoundClassNP	180
NamedEntityCompound	145
CompoundTransitiveVP	70
CompoundAssociativeVPNoTense	46
CompoundAtomicVPWithCirc-NoTenseNoMood	28
CompoundAttrGroup	24
VerbCoordCompoundNoTense-NoMood	16
CompoundAssociativeObject	12
VerbCoordCompoundNoTense	3
CompoundAssociativeVPNoTense-NoMood	2
CompoundAssociativeVP	2
VerbCoordCompound	1
CompoundTransitiveVPNoTense-NoMood	1
CompoundAtomicVPWithCirc-NoTense	1

Table 2: Statistics on nominal compounds in the ontology.

Compound class	Absolute frequency
NPGenCompound	248
CompoundClassNP	26
CompoundNRootGroup	24
AdjunctCompound	15
NamedEntityCompound	8
CompoundAttrGroup	4
VerbCoordCompound	2
VerbCoordCompoundNoTense	1

Table 3: Statistics on verbal compounds in the ontology.

Compound class	Absolute frequency
CompoundVPWithCirc	44
CompoundTransitiveVP	19
VerbCoordCompoundNoTense-NoMood	6
CompoundAssociativeVPNoTense	3
CompoundAssociativeVP	1

7 CONCLUSIONS AND FURTHER WORK

The current results of the formal grammatical and ontological modeling of Tibetan compounds presented in this article represent the first of its kind consistent systematic formal description of this material, which is confirmed by corpus data. This description does not claim to be universal for the entire Tibetan language, but it not only covers all types of Tibetan compounds that researchers have introduced before, but also includes models of classes of compounds that have not been previously described. Moreover, this model is part of the Tibetan language module of a working automatic text processing system, and it is verified by analyzing the results of the automatic syntactic and semantic annotation of the corpus of texts. This model still does not cover all cases of compounds use in the corpus, namely, some types of contractions found in poetic works. An exhaustive modeling of such phenomena is planned to be performed within the framework of this study.

ACKNOWLEDGMENT

This work was supported by the Russian Foundation for Basic Research, Grant No. 19-012-00616 Semantic interpreter of texts in the Tibetan language.

REFERENCES

The Basic Corpus of the Tibetan Classical Language [Online]. 2019. Available at: http://corpora.spbu.ru/bonito/index_gram.html. Accessed at: 19 May 2019.

The Corpus of Indigenous Tibetan Grammar Treatises [Online]. 2019. Available at: http://corpora.spbu.ru/bonito/index.html. Accessed at: 19 May 2019.

Great Encyclopedical Dictionary [Bolshoy entsiklopedicheskiy slovar]. Linguistics [YAzyikoznanie]. Scientific Publishing House “Great Russian Encyclopedia” [Nauchnoe izdatelstvo «Bolshaya Rossiyskaya entsiklopediya»], Moscow, 1998, 2nd (reprint) of linguistic encyclopedic dictionary edition.

Beyer, S. (1992). *The Classical Tibetan Language*. State University of New York, New York.

Dobrov, A. (2014). Semantic and ontological relations in aiire natural language processor. In *Semantic and ontological relations in AIIRE natural language processor*, pages 215–222, Rzeszow-Sofia. ITHEA.

Dobrov, A., Dobrova, A., Grokhovskiy, P., Smirnova, M., and Soms, N. (2018a). Computer ontology of tibetan for morphosyntactic disambiguation. In Alexandrov, D. A., Boukhanovsky, A. V., Chugunov, A. V., Kabanov, Y., and Koltsova, O., editors, *Digital Transformation and Global Society*, pages 336–349, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-02846-6_27.

Dobrov, A., Dobrova, A., Grokhovskiy, P., Smirnova, M., and Soms, N. (2018b). Modeling in a computer ontology as a morphosyntactic disambiguation strategy. In Sojka, P., Horák, A., Kopeček, I., and K., P., editors, *Text, Speech, and Dialogue. TSD 2018. Lecture Notes in Computer Science*, volume 11107, pages 76–83, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-00794-2_8.

Dobrov, A., Dobrova, A., Grokhovskiy, P., and Soms, N. (2017). Morphosyntactic parser and textual corpora: Processing uncommon phenomena of tibetan language. In *Proceedings of the International Conference IMS*, pages 143–153, Saint-Petersburg. https://doi.org/10.1145/3143699.3143719.

Dobrov, A., Dobrova, A., Grokhovskiy, P., Soms, N., and Zakharov, V. (2016). Morphosyntactic analyzer for the tibetan language: aspects of structural ambiguity. In *International Conference on Text, Speech, and Dialogue*, pages 215–222. https://doi.org/10.1007/978-3-319-45510-5_25.

Grokhovskii, P. and Smirnova, M. (2017). Principles of tibetan compounds processing in lexical database. In *Proceedings of the International Conference IMS*, pages 135–142. SCITEPRESS. ISBN: 978-1-4503-5437-0 DOI 10.1145/3143699.3143718.

Grokhovskii, P., Zakharov, V., Smirnova, M., and Khokhlova, M. (2015). The corpus of tibetan grammatical works. *Automatic documentation and mathematical linguistics*, 49(5):182–191. https://doi.org/10.3103/S0005105515050064.

Gruber, T. (1993). *A Translation Approach to Portable Ontology Specifications*, volume 5 of *Knowledge Acquisition*. Stanford University. Computer Science Department. Knowledge Systems Laboratory. https://doi.org/10.1006/knac.1993.1008.

Leont’eva, N. (2003). Ruslan semantic dictionary as a tool for computer understanding [semanticheskij slovar ruslan kak instrument kompyuternogo ponimaniya]. In *Understanding in communication. Proceedings of the scientific practical conference [Ponimanie v kommunikacii. Materialy nauchnopraktich-eskoj konferencii]*, pages 41–46, Moscow.

Maslov, Y. S. (1998). Verb [glagol]. In Yartsev, V. and N.D, A., editors, *Great Encyclopedical Dictionary [Bolshoy entsiklopedicheskiy slovar]: Linguistics [YAzyikoznanie]*. Great Russian Encyclopedia [Bolshaya Rossiyskaya Entsiklopediya].

Matuszek, C., Cabral, J., Witbrock, M. J., and DeOliveira, J. (2006). An introduction to the syntax and content of cyc. In *AAAI Spring Symposium: Formalizing and Compiling Back-ground Knowledge and Its Applications to Knowledge Representation and Question Answering*, pages 44–49.

Melcuk, I. (1995). *Phrasemes in language and phraseology in linguistics*. Lawrence Erlbaum, New Jersey.

Mel'čuk, I. and Žolkovskij, A. (1984). *Explanatory Combinatorial Dictionary of Modern Russian*. Sonderband. Ges. zur Förderung Slawistischer Studien.

Pelletier, F. (1994). The principle of semantic compositionality. *Topoi*, 13(11).

Rubashkin, V. S., Fadeeva, M. V., and Chuprin, B. Y. (2012). The technology of importing fragments from owl and kif-ontologies [tekhnologiya importa fragmentov iz owl i kif-ontologij]. In *Proceedings of the conference "Internet and modern society [Materialy nauchnoj konferencii "Internet i sovremennoe obshchestvo"]*, pages 217–230.