


A Study of the Tibetan Linguistic Picture of the World Using Computer Ontology¹

Maria Smirnova

(Saint Petersburg University)

he research presented in this article is a summary of several research projects aimed at the creation of a full-scale natural language processing engine based on a consistent formal model of Tibetan vocabulary, grammar, and semantics, verified by and developed on the basis of a representative, hand-tested corpus of texts.

The *Basic Corpus of Classical Tibetan*² and the *Corpus of Indigenous Tibetan Grammar Treatises*³ comprise 34,000 and 48,000 tokens,⁴ respectively. Tibetan texts are represented both in the Tibetan Unicode script and in standard Wylie romanization.⁵ These corpora are developed, annotated, and tested manually by Tibetologists, and in this sense, are unique.

The ultimate goal of our project is to create a formal model (a grammar and a linguistic ontology) of the Tibetan language, including morphosyntax, syntax of phrases, hyperphrase unities,⁶ and semantics, that can produce a correct morpho-syntactic, syntactic, and semantic annotation of the corpora without any further manual corrections.

This study is based on the technologies and tools of the AIIRE project.⁷ AIIRE⁸ is a free open-source natural language understanding

¹ This work was supported by the Russian Foundation for Basic Research, Grant No. 19-012-00616, "Semantic Interpreter of Texts in the Tibetan Language."

² "The Basic Corpus of the Tibetan Classical Language." http://corpora.spbu.ru/bonito/index_gram.html. Accessed March 3, 2020.

³ "The Corpus of Indigenous Tibetan Grammar Treatises." <http://corpora.spbu.ru/bonito/index.html>. Accessed March 3, 2020.

⁴ A token is the smallest unit that divides each corpus. Its usual application refers to lexical tokens, words or other atomic parse elements.

⁵ Grokhovskii *et al.* 2015: 182–191.

⁶ A hyper phrase is a segment of text in the form of a sequence of two or more independent sentences, united by a common theme in semantic blocks.

⁷ Dobrov *et al.* 2016: 215–222.

⁸ "AIIRE—Artificial Intelligence Information Retrieval Engine." <http://svn.aiire.org/repos/tproc/trunk/t/>. Accessed March 3, 2020.

system (NLU system).⁹ This system implements a full-scale procedure for natural language understanding—from graphematics¹⁰ to morphological annotation and syntactic parsing—and even includes semantic analysis.

The development of the morphosyntactic analyzer of Tibetan texts and the formal grammar necessary for it has been very complicated due to the ambiguity of both the segmentation of Tibetan texts into morphemes (since there are no word delimiters between word forms in Tibetan writing) and the syntactic parsing. Syntactic parsing without the help of semantic restrictions leads to combinatorial explosions. To resolve the problem of morphosyntactic ambiguity, our team¹¹ decided to use the AIIRE tool for semantic analysis—thus developing a *computer-based linguistic ontology*.

Such linguistic ontologies are designed for automatic processing of unstructured texts. Units of linguistic ontologies are based on meanings of actual natural language phrases. Ontologies of this kind model a linguistic picture of the world represented by language semantics.

The present article describes the process of developing a computer ontology of the Tibetan language, as well as a methodology of its practical application for the current corpus and opportunities for its use in the field of interdisciplinary studies of Tibetan linguistics and wider research on the Tibetan linguistic picture of the world.

⁹ AIIRE is developed and distributed under the terms of the GNU General Public License that is a free, copyleft license for software and other kinds of works. The name of the system, GNU, is a recursive acronym meaning “GNU’s Not Unix.” This acronym is a way of both paying tribute to the technical ideas of Unix and saying that GNU is something different. Technically, GNU is like the operating system Unix. But unlike Unix, GNU gives its users freedom.

¹⁰ The initial stage of text processing usually includes the segmentation of input text into graphemes and further refers to the recognition of words and additional graphemic components (e.g., punctuation marks). For inflectional languages the input units are easy to identify as word forms, separated by space, punctuation marks, etc. However, that is not the case with the Tibetan language, as there are no word delimiters. The AIIRE system performs the segmentation of the input string into elementary units by using a special algorithm (Aho-Corasick) that allows for the detection of all possible substrings of the input string according with a given dictionary. See Dobrov *et al.* 2017 for more details on the implementation of this approach.

¹¹ The linguistic ontology presented in this research is the result of joint efforts of an entire team including Aleksei Dobrov, Anastasia Dobrova, Olga Dzhangolskaya, Yana Khramova, Anna Kramskova, Ksenia Rastorgueva, Nikolay Soms, and Viktor Zakharov. This work would not have been possible without its founding member and team leader of all our projects—Pavel L. Grokhovskiy who passed away on December 17, 2018.

1. Related Work

Computer linguistics generally defines “ontology” as “an explicit specification of a conceptualization.” This definition was popularized by Thomas R. Gruber, where conceptualization is “an abstract, simplified view of the world that we wish to represent for some purpose.”¹²

Without claiming any changes to this *de facto* standard, our team has to clarify that as researchers we do not mean just any “specification of a conceptualization” by this term, but rather a computer ontology, which is defined as a database consisting of concepts and relations between them. 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 (i.e., between two concepts) and directed. Each relation is directed from the subject of the relation to the object. For example, the relation “to have a part of the body” should be directed from its subject (the concept “any creature”) towards its object (a concept, denoting “any part of the body”). In turn, the concept “leg,” for example, is the subject of the reverse relation—“to be a part of somebody’s body.”

Linguistic ontologies are designed for automatic processing of unstructured texts.¹³ Units of linguistic ontologies are based on meanings of real natural language expressions.¹⁴

In the first generations of natural language understanding systems (NLU systems), ontologies were used as semantic dictionaries. In the early 1990s, several scholars already used the term “ontology” in the most general sense, which allowed linguistic thesauri to be considered as types of ontologies. The WordNet computer thesaurus has come to be called an “ontology,” and this trend has only been growing in the majority of modern works.

Thesauri, including the WordNet, reflect more or less specified semantic relations between lexical units (words): synonymy, hyponymy, hypernymy, antonymy, meronymy, holonymy, logical entailment, the relation of an adjective to a noun, etc. These relations can be used to perform lexical disambiguation. Unfortunately, these relations alone are not enough to solve the problem of lexical or morphosyntactic ambiguity, especially in Tibetan, since they do not reflect semantic valencies.

¹² Gruber 1993: 199.

¹³ Unstructured data is the information that either does not have a pre-defined data model or the one which is not organized in a pre-defined manner. A text is considered unstructured data.

¹⁴ Dobrov *et al.* 2018: 340.

In contrast, the *linguistic* ontologies model strictly specified relations between concepts such as the relation between a physical object and its parts (meronymy); the relations between the agent and the actions that the agent can perform; the relations between an action and objects towards which this action can be directed, etc. Some of these concepts represent meanings of different lexical units, others have no representation in vocabulary, but are necessary for its modeling.¹⁵ This difference between thesauri and linguistic ontologies becomes obvious in the attempt to create inference systems: linguistic ontologies are built on the basis of logical formalisms and corresponding inference rules. In contrast, thesauri generally do not provide any native mechanisms for logical inference.¹⁶ A semantic dictionary is a description, whereas an ontology is a model that predicts and explains this description and can be used and developed with much higher efficiency.

Ontologies are used for various tasks in natural language processing systems: from primitive problems of named entity recognition or text classification¹⁷—for which, to a certain extent, thesauri can also be used—to tasks of full-scale semantic analysis of texts, which involves inference of meanings based on individual lexical units. Ontologies can also be used for tasks that require syntactic and lexical disambiguation based on strict semantic relations that thesauri do not provide. Such relations include those between classes of entities and actions that these entities can perform or the states in which they can be; relations between these actions and states, on the one hand, and their objects, on the other; relations between actions or states and their objects; all kinds of relations that can be expressed by the genitive case—the relation between an object and its owner, between a part and the whole, or the most complicated relationships between people, organizations, and societies, expressed by the genitive construction—as well as all kinds of relations expressed by prepositions, etc.¹⁸

Until now, the task of creating a universal linguistic ontology has been set only for European languages, which resulted in the spread of a number of incorrect assumptions concerning ontological semantics in general. It was suggested by some researchers that a universal computer ontology may not depend on a particular language. But, in fact, this universality falters due to the specificity of

¹⁵ Dobrov 2014: 151.

¹⁶ Dobrov *et al.* 2018: 339.

¹⁷ Sánchez-Pi, Martí, and Garcia 2016: 48–58; Zhou and El-Gohary 2015; Sánchez-Cisneros and Aparicio 2013: 622–627; Lytvyn *et al.* 2017: 229–240; Abdollahi *et al.* 2019.

¹⁸ Kang and Lee 2001: 199–220; Jensen and Nilsson 2006: 229–244; Dobrov 2014.

each language. Ontology is universal in the way that it concerns diverse subject matters, but not in regard to many languages. It is obvious that an ontology cannot possibly be fully independent of a specific language because not all linguistic units of those languages have a direct analogy in other languages. Different general concepts in different languages have specific linguistic units that have individual semantic meanings not represented in every language of the world. Furthermore, the structuring of the world itself and thus its concepts and lexical meanings may differ significantly from language to language, which has an effect not only on lexical but also on grammatical semantics.

Even though scholars are working on the tools for processing Tibetan texts in different countries (e.g., Germany, Great Britain, China, USA, Japan, Netherlands), there is still no conventional standard of corpus annotation for Tibetan language material. A number of recent studies were primarily aimed at developing solutions for the initial stages of Tibetan NLP, such as word segmentation and part-of-speech tagging. No attempts have been made to develop the Tibetan thesaurus, let alone ontology for the entire Tibetan language.

2. The Structure of the Computer Ontology

Our Tibetan ontology is developed within the framework of the AIIRE ontology editor software.¹⁹ In the AIIRE project, “ontology” is understood as a consistent classification of concepts that unite the meanings of Tibetan linguistic units, including morphemes and idiomatic morphemic complexes.

Concepts are interconnected with different semantic relations. To create a new concept, it is compulsory to incorporate this concept into the general classification hierarchy according to class-superclass relations (hypo/hyponymy). Therefore, the whole ontology denotes one common superclass.

The ontology models the meanings of atomic linguistic units (morphemes) and of idiomatic combinations of these units, including nominal and verbal compounds, idiomatic nominal groups, as well as adjectival and adverbial groups. In all these cases, in addition to the meanings of each idiomatic expression, meanings of its

¹⁹ The ontology itself is available at the AIIRE website in a snapshot (<http://svn.aiire.org/repos/tibet/trunk/aiire/lang/ontology/concepts.xml>; accessed March 14, 2020) and it is also available for unauthorized view or even editing (please refer to <http://ontotibet.aiire.org>; accessed March 14, 2020). The editing permission can be obtained through an access request.

components are also modeled in the ontology so that they can be interpreted in their literal meanings as well.

To model a new concept, a researcher needs to create an expression entry in the ontology. An expression is analog to a heading word in a dictionary entry (e.g., the expression *sbrang rtsi* in Fig. 1). Then, a researcher gives the meaning of the expression and provides a translation and description (or interpretation) of the expression in Russian.²⁰ These entries are intended to facilitate a common understanding of the decisions made by project participants in the process of editing the ontology (the choice of hypernym, the establishment of certain semantic relations, etc.). The main source for establishing the basic meaning of each expression is a text or texts in the employed corpora where the expression is used.



<p>1. honey; a sweet, viscous food substance produced by bees and some related insects: [2032] 1) <i>bung bas bsags pa'i rtsi mngar mo/ ... spu gri'i so la sbrang rtsi byug pa/ ... ming gi rnam grangs la mngar ldan dang/ spra tshil lo/ 2) ro mngar/ zhu rjes drod/ nus pas lus stobs nyams pa gso ba dang/ mig nad bar 'grib/ rtsa dkar gyi nad/ dbang po mi gsal ba/ sha 'phel gyi nad bcas la phan/ bad kan dang/ chu ser la phan pa'i sman gyi nus pa rtsar 'khrid par byed/</i></p>
<p>Token type: noun phrase with genitive compound</p>
<p>synonyms</p>
<p>hypernyms</p>
<p>1. མཚོ (substance that influences color and taste...) —</p>
<p>hyponym: 4</p>
<p>hyponyms</p>

Fig. 1 — The expression *sbrang rtsi* in the computer ontology

Each expression, the meaning of which is modeled in the ontology, is also provided with a full-scale interpretation in Tibetan from *The Great Tibetan-Chinese Dictionary* (see Fig. 1).²¹ If according to the dictionary, the expression has several meanings, then the one used in the particular context of the corpus is translated into Russian (except for the case when the expression is defined in the dictionary through synonyms). For each concept, a separate type of token is established.

²⁰ The Russian language is the language of the software interface, including the ontology itself. In the ontology, Russian is also used for technical classes and to describe verbal semantics and relations between concepts.

²¹ Zhang 1985: 2032.

The number of token types in the ontology has been continuously expanded: with the development of the formal grammar, new types of tokens were added into the ontology. For example, new types of nominal and verbal compounds were identified.

The researcher establishes different relations between concepts. The relation of synonymy is always absolute, which suggests complete correspondence of referents with possible differences in significations. In linguistics, synonyms are usually defined as words that are close in meaning. In the computer ontology, synonyms are meanings of different linguistic units that have strictly identical denotations.

Concepts form synonymic sets. Each element of the set has the same attributes, i.e., the same relations and objects of these relations. The variance of significations within a synonymic set is compensated by automated logic rules: if Y is the synonym of X, then X has the same attributes as Y, and Y has the same characteristics as X. For example, if the concept *deb* “book” is the subject of the relation “to have been written by an author,” its synonym *dpe cha* “book,” is also considered to be the subject of the same relation. In other words, anything that could cogently be said about a *deb* should also apply to a *dpe cha*.

Hypo-hypernymy is established between classes and subclasses or between classes and instances when one concept (hyponym) is a token of another (hypernym). For example, the class *pho gsar* “young man,” is a subclass of *pho* “man,” which is a subclass of *mi* “human being.” If there is a lacuna in Tibetan, it is possible to use a Russian hypernym. For example, the hierarchy of hypernyms for the Tibetan concept *lag pa* “hand” is presented in Fig. 2.

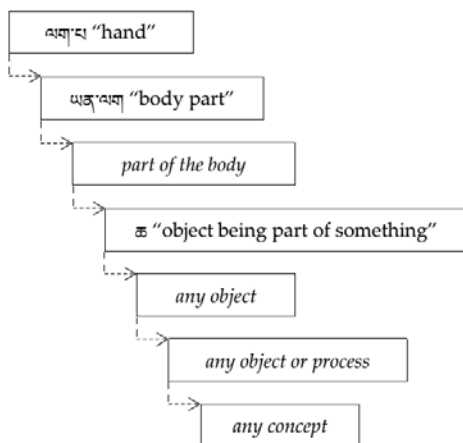


Fig. 2 — Concept hierarchy of the expression *lag pa* in the computer ontology

Here, the conceptual gaps in the hierarchy are rendered in English (originally, they are in Russian, as it is the default language of the ontology interface). If, during the development of the ontology, some gaps remain, it means that the concept hierarchy probably needs additional manual correction.

Each concept must have at least one hypernym, except for the ontology root concept.²² In order to determine semantic valencies, it is necessary to create a concept hierarchy that has basic classes—for example, the class “person.” Basic classes usually have a large number of relations, which appear in genitive constructions, verb valencies (a number of predicate arguments), etc.

Modeling verb meanings in the ontology is made with the use of special tools that allow speeding up and partial automating of verbal concept modeling. The AIIRE *Ontohelper* is used together with the main AIIRE ontology editor web interface to build a complete hierarchy of superclasses for any verb meaning in the ontology.

The logic behind the *Ontohelper* is also based on the division of verbs into dynamic (terminative and non-terminative) and static ones.²³ Dynamic verbs express actions, events, and processes associated with different changes. Static verbs express states, relations, or qualities.²⁴ A terminative verb denotes an action that has a limit in its development. A non-terminative verb denotes an action which does not admit any limit in its development. For example, one can take the verb “to sing.” In the sentence, “she sang,” the verb is non-terminative since the duration of her singing is not defined. However, in the sentence “she sang a song,” the verb is terminative, as it is clear that she was singing for the precise amount of time limited by the duration of that song. The verb can also be defined as terminative just by the meaning of its root.

When using the *Ontohelper* editor, it is necessary to determine whether the verb being modeled denotes an action, activity, or state. Terminative, non-terminative, and static verb meanings correspond in the ontology to subclasses of concepts “to perform an action,” “to perform an activity,” and “to be in a state,” respectively. The editor of the ontology indicates the basic class for subjects of the verb to be modeled, 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, marked in Tibetan with

²² “Root concept” is a single concept, the common superclass of the entire ontology. It is named as “any concept” in the ontology.

²³ Maslov 1998.

²⁴ Ibid: 105.

dang).²⁵ As a result, the ontological editor builds a complete class hierarchy for the modeled verb meaning. For example, modeling the verb *sbyin* “to give,” requires the creation of 523 classes of verb concepts, including its direct hypernym “to perform an action by any creature directed to any object addressed to any creature.” If the given hierarchy includes classes that already exist in the ontology, they are not rebuilt.

Within the framework of the present research, 4335 concepts, including 3943 meanings of Tibetan expressions, were modeled in the ontology.

3. Methodology of the Ontology Implication

Basically, we work with four types of errors: unrecognized units, combinatorial explosions, breaks in syntactic trees and overlaps thereof. Unrecognized fragments are the fragments that the ontology cannot parse. Combinatorial explosions are cases of exponential growth in the number of possible parsings. As the length of the parsed text, and, thus, the number of its ambiguous fragments, increase, parsing permutations increase as well. Syntactic trees describe a method of formulating a hierarchy of the syntactical relationship between expressions in a sentence, each belonging to parts of speech, to noun or verbal phrases, up to the level of the sentence itself. Breaks in these trees occur when the ontology fails to fully map these nested relations for an expression up to the level of the sentence. Overlaps occur when fragments of text belong to two syntactic trees, but neither of the trees completely covers the text to which they belong.

We use our ontology for the consistent elimination of these annotation errors, starting with the most important and frequent ones. As said above, the main reason for the use of the ontology was the need to perform morpho-syntactic disambiguation. This includes dealing with a special type of annotation errors—combinatorial explosions. Most combinatorial explosions were caused by the prevalent use of idiomatic morphocomplexes and compounds in the Tibetan language. Thus, in the initial stage, the computer ontology was used to model the meanings of Tibetan nominal and verbal compounds found in the corpus texts. The work was carried out simultaneously with all the texts of the corpus.

The result of this work was the classification of Tibetan compounds. The classification not only covers all types of Tibetan

²⁵ Dobrov *et al.* 2019: 147.

compounds that researchers have introduced before, but also includes models of classes of compounds that have not been previously described. Different types of compounds require the introduction of different semantic relations between its components in the computer ontology.

For example, Tibetan often combines letters or exponents of arbitrary Tibetan morphemes with a noun root, e.g., *la sgra*, which denotes “grammatical marker *la*.”²⁶ This class was called the “named entity compound” and was introduced into the formal grammar. The *la sgra* class is a subclass of named-entity nomination, where the name of the entity is a letter or an exponent of any Tibetan morpheme, in this case, *la*. To ensure the correct parsing of the compound *la sgra*, it is necessary to connect the expressions “linguistic unit” (that is the basic class of *sgra* “grammatical marker”) and “any exponent” (the basic class for all exponents of any Tibetan morpheme) with the relation “to denote a concept” in the computer ontology.²⁷

The next step in resolving morpho-syntactic ambiguity was the establishment of the following types of restrictions in the computer ontology: the restriction on adjuncts, the restriction on genitive relation, the restriction on classes of direct objects and subjects of verbs.

Restrictions on the general genitive relation “to have any object or process (about any object or process)” are imposed by establishing specific relation subclasses between basic classes in the ontology. For example, to exclude the possibility of the first version of parsing (1.1) in the example (1) below, the concept *lus* “physical body,” was allowed to possess a genitive relation of “to have a body (about human being)” only when connected with the concept *mi* “human being.” This facilitated the exclusion of the version of parsing in which “fame” can have a body.

- (1) བླ་མཁའ་པོ་ལྷོ་ལྷོ་
grags-pa 'i lus
 be_well-known-NMLZ GEN body
 (1.1) ‘body of fame’

²⁶ *La sgra* and *la don* are paired terms of the Tibetan linguistic tradition. The first denotes the form, the second—the meaning (i.e., “grammatical markers with meaning of *la*”). This is a typical opposition for the Tibetan linguistics (for example, *sgra'i sbyor tshul* denotes “the way of joining [grammatical marker] according to its form” (i.e., the rules for choosing allomorphs) while *don gyi 'jug tshul* is “the way of joining [grammatical marker] according to its meaning.”

²⁷ Dobrov *et al.* 2019: 149.

(1.2.) 'body of a famous [person]'²⁸

At the moment, the concept *mi* "human being," has the following genitive relations in the computer ontology:

- to be a person whose activity is related to a thing;
- to be a person born or living in any place;
- to be a person occupying a certain position;
- to have a text;
- to be a participant of a social process;
- to be a person whose activity is related to animals;
- to have age (i.e., to be of a certain age);
- to have a body.

According to the inheritance of the attribute's rule, all following concepts that are hyponyms of the class *mi* inherit its semantic relations: any person of a certain ethnicity; any person of a certain belief; any person of a certain age; any person of a certain occupation; any person with a certain ability; any person of high social status; any person engaged in a certain activity at a certain level; any person related to some institution; any person with a certain skill; any person who was born or lived in a certain place; any person in a certain relationship with other people; any person characterized by a certain social connection with other people; *khyim bdag* "housekeeper"; *grong pa* "neighbor"; *chos pa* "religious devotee"; *dgra* "enemy"; *dpa' bo* "hero"; *pha rol po* "opponent"; *pho* "male"; *btsun pa* "religious teacher"; *bud med* "woman"; *mkhas mchog* "supreme sage"; *mtha' khob* "barbarian"; *mu to pa* "poor man"; *mo* "female"; etc.

Tibetan adjuncts are placed after the noun they modify. Due to the absence of word delimiters (spaces) in the Tibetan writing system, adjuncts cannot be graphically distinguished from elements of a compound, and in this way may cause incorrect parsing. For example, the compound (2) in the example below may be misinterpreted either as "father-mother" ("a father, who is also a mother")—*ma* "mother," is interpreted as an adjunct—or as "father's mother" (a noun phrase with a genitive compound). However, the only correct interpretation is "father and mother" (a noun root group compound). While the second interpretation (which is, moreover,

²⁸ Hereinafter the interlinear morpheme-by-morpheme glosses are made according to the Leipzig Glossing Rules. <https://www.eva.mpg.de/lingua/resources/glossing-rules.php>. Accessed May 20, 2020.

logically possible) can be eliminated by just setting the correct token type in the ontology (because both the second and the third interpretations imply that the phrase is a compound), the first interpretation cannot be eliminated in this way.

(2) བཤམ
pha ma
 father_mother
 'father and mother'

Thus, only semantic restrictions can eliminate semantically incorrect adjunct versions of parsings. This is achieved by limiting possible equivalence relations ("to be equivalent to an object or process"). Basic classes were connected with themselves to this relation so that only concepts that inherit these classes could be interpreted as adjuncts for each other. Though, as in the case of *pha ma* this is not necessarily the case.

Restrictions on subjects and direct objects of verbs were necessary for the correct analysis of compounds and idioms, as well as for eliminating unnecessary versions of syntactic parsing. As it was said before, these restrictions are imposed by specifying the correct classes of verb subjects or objects in the *Ontohelper* editor, so that only concepts that inherit these classes can perform or be the object of the action denoted by a certain verb.

At the present stage, we continue developing the ontology and using it to eliminate annotation errors. The ultimate goal of our project is to create a complete semantic annotation of all texts in the corpus.

4. Study of the Lexical Semantics Using the Computer Ontology

The specific nature of the Tibetan linguistic picture of the world is particularly evident in the field of lexical semantics. When we model concepts in the computer ontology, we need to classify them. Since our corpus mainly includes Tibetan grammar texts and texts on the theory of writing, these classifications are specific. We model all necessary semantic relations in order to perform correct semantic parsing of a given text. This helps reveal the most frequent terms, lingua-specific concepts, typical semantic relations, etc. Thus, the creation of the computer ontology discovers features of the Tibetan picture of the world (notably, the scientific picture of the world described in section 5 of the present article) and itself becomes a formal model of the Tibetan linguistic picture of the world.

Buddhism, to a great extent, determines the characteristics of the Tibetan linguistic picture of the world. Terms in the Buddhist doctrine are basic lingua-specific concepts and are key to Tibetan culture. When modeling such concepts in the computer ontology, a Tibetologist usually encounters a number of difficulties. In order to create interpretations or identify hypernyms for some specific concepts, one could rely on the context in which a term is used, as well as on the definition of this term in a Tibetan definition dictionary. In cases when the term is a Buddhist compound, formulating its interpretation and identifying its hypernym can be facilitated by the term's internal structure. In particular, the analysis of Tibetan nominal and verbal compounds utilizing the computer ontology made it possible to identify typical syntactic and semantic structures.²⁹ For example, according to the syntactic structure, the compound *sbrang rtsi* in Fig. 1 is a noun phrase that is a genitive compound. According to the semantic structure, it refers to subordinate compounds in which the second element (*rtsi* "substance") is a hypernym of the whole compound.

For the Tibetan picture of the world, the difference between animals and humans, and the difference between living beings with a dualistic mind trapped in cyclic existence and the Buddha is relevant. These features require a construction of several complex hierarchies of concepts in the computer ontology. Thus, in the Tibetan linguistic picture of the world, the most frequently used class of subjects of verbs is the basic class "any creature" (while for the Russian language it is the class "any person"). At the moment, this basic class includes several hyponyms in the ontology, some of which include only humans (e.g., *mi* "human") and others that unite people and animals (e.g., *sems can* "sentient being who has a dualistic mind"; *'gro ba* "migrator"; and *skyes ldan* "having a birth"), or others that even unite people, gods and Buddhas (e.g., "any creature that is not an animal").

The relation between Buddhism and the linguistic picture of the world also appears in fundamental categories. Thus, Buddhist discourse is highly parameterized by descriptions of space—one of the basic categories—and distinguishes between proximity and distality. In other linguistic pictures of the world, opposition-primitives such as "top/down," "forward/reverse," etc., are used to describe proximal space that directly "adjoins" a person. The basic meaning of these oppositions is relative to the human body. Distal space parameters, on the other hand, are associated with

²⁹ For more details on the semantic types of Tibetan compounds, see Grokhovskii and Smirnova 2017. For different types of their syntactic structure, see Dobrov 2018.

“hostility/friendliness” of corresponding objects. Therefore, the most significant opposition for the distal space is “friend/foe.” Thus, the distal space is defined via the person *cum* “social organism.”³⁰

But in the Tibetan linguistic picture of the world, distal space becomes metaphoric for the object’s relationship with Buddhist religious doctrine. Rather than placing objects on a continuum from social inclusion to exclusion, Tibetan distal words describe another continuum with the Buddha and his teaching as the starting point.

For example, the Tibetan compound *mtha’ khob* can denote any “borderland” or “suburb,” as well as “barbarian lands, unfamiliar with the higher culture of Buddha’s teaching,” or even a “person who does not practice Buddhism, or who does not belong to the Buddhist spiritual community.” The connection of spatial metaphors with Buddhism is also demonstrated by the opposition between the Tibetan terms *nang pa* “Buddhist,” and *phyi rol pa* “non-Buddhist,” literally meaning “insider” and “outsider.”

Thus, the computer ontology, which includes different classification hierarchies of concepts in the Tibetan language, can be used as a model and tool for studying the Tibetan linguistic picture of the world as it is given in Tibetan texts.

5. Modeling Tibetan Concepts Related to Subject Areas of Knowledge

Since the Tibetan-Russian corpus includes texts on the traditional Tibetan sciences, linguistic works (*sgra’i rig pa*) constitute one of its major genres. Thus, a large number of modeled concepts refer to grammatical terms³¹ and special lexis of the theory of writing. Therefore, they reflect the Tibetan scientific picture of the world. Tibetan linguistics is mainly based on grammars created by Buddhist scholars and is strongly connected with the Indian tradition. Moreover, Buddhist lingua-specific concepts and ideas were widely used in texts on other sciences. It is also typical for different traditional Tibetan disciplines to use the same terms with different meanings.

³⁰ Kasevich 1996: 133–134.

³¹ It should be noted that the word “term” in its strict modern definition does not fully apply to the special lexis of the Tibetan fields of medieval knowledge, including the Tibetan grammatical tradition. Tibetan grammar terms do not meet all the criteria of a scientific term (such as, for example, monosemy or motivation, that is the term itself having such sufficient semantic transparency that an approximate understanding of the concept denoted by a term can be formed). In this situation, it is more appropriate to talk about pre-terms—lexical units used as terms in subject areas for naming newly formed concepts, but not meeting the basic requirements of a scientific term (Grinev-Grinevich 2008: 44).

In the Tibetan grammar texts created within the framework of the Buddhist religious tradition, the morpheme *dbyangs* “voice/sound” was used to mean “vowel phoneme.” In contrast, in the treatises on music, the same term denotes “melody,” or even functions as an element in the compound name of the bodhisattva Mañjuḥṣa (‘Jam dbyangs; literally “tender melody”).

Apart from indigenous linguistic terms, religious and philosophical terms were also widely used in Tibetan grammar treatises. When using the computer ontology, the researcher can indicate a subject area for concepts of the Tibetan traditional sciences, be it linguistics, Buddhist religious doctrine, etc. This helps the study of Tibetan terminological fields, their structure, the interconnection of terms, terminological polysemy, homonymy, etc.

For example, in the texts of our corpus, in addition to grammatical terms, general scientific terms are also common. Most of them denote various text parts or sections. Thus, at the moment, the expression “text structural unit” in the ontology already has 14 hyponyms: *nang gses* “text sub-division”; *mtha’ dpyod* “thorough study”; *re’u mig* “table”; *sa bcaḍ* “text section”; *skabs don* “text section that reveals the main theme”; *sdom tshig* “concise conclusion”; *bam po* “section or chapter of the text”; *skabs* “chapter”; *sdom* “conclusion, summary”; *mchod brjod* “expression of worship for the Buddha and the gods”; *nang tshan* “text section”; *mjug bsdu* “summary of a message”; *mdor bstan* “synopsis”; and *’gyur phyag* “homage to the translator.”

A number of concepts denote various types of scientific, literary, or religious texts. In particular, the Tibetan scientific tradition identifies basic texts as *gzhung* and their numerous commentaries—*rgyas bshad* “detailed commentary”; *rnam bshad* “thorough commentary”; etc.

A large number of concepts represent the class *don* “meaning [of the text],” common to all Indo-Tibetan traditional sciences, which includes concept-hyponyms important to Tibetan scientific literature, such as *gnad don* “key meaning”; *gzhung don* “the main meaning of the text”; *dgongs don* “implied meaning”; *brjod don* “topic”; *go don* “core meaning”; etc.

There were common ways of term formation in the Tibetan scientific tradition such as terminologization of common words, compounding, and borrowing. In some cases, nominalized verb forms acquired narrow terminological meanings. These forms include nominals produced by the syllabic formative, nominalizer *-pa*, as well as the forms formed by adding nominalizing suffixes with the meanings “method,” “place,” “path,” etc. to the verb. The unfinished state of the terminologization process, the closeness of the Tibetan special lexis to common language, the presence of a large number of

consubstantial terms—found both in everyday speech and professional terminology³²—often lead to hypo-hyperonymy relations between concepts of a single expression. For example, in the grammatical terminological field, the Tibetan *rtags* is “a grammatical sign which marks various grammatical meanings.” At the same time, as a common word, it means “sign, tool for transmitting and receiving information, localized in any space.” Thus, its common meaning is a hypernym of its meaning as a grammar term.

In those cases when certain disciplines employ idiosyncratic hypo-hypernym relationships, or when the semantic valences of these terms differ from normal usage, the ontology used the relation “to have a typical representative (about the class of objects)” and the inverse relation, that is “to be a typical representative of the class.” Since, in the case of *rtags*, the class and the typical representative are expressed by one word in Tibetan, the expression denoting a typical representative was indicated in the ontology in Russian. In this way, we create a separate expression “a grammatical sign” and connect it with the relation “to be a typical representative of” with the expression *rtags*.

A number of Tibetan terms are formed by adding a numeral to a noun, thus denoting a collection of objects (for example, *dus gsum* “three verb tenses”; *dus bzhi* “four seasons”; *byung ba lnga* “five elements”; etc.). To connect collections and their elements in the computer ontology, the relation “to include objects of a class” (and the inverse, “to be an object of a class”) were used. For example, the term *dus bzhi* is connected through this relation with the concepts *dgun kha* “winter”; *dpyid ka* “spring”; *dbyar kha* “summer”; and *ston kha* “autumn.”

Polysemy and the absence of unique meanings of morphemes are typical features of Tibetan terminology in general and grammatical terms in particular. Not all contexts can reveal the particular meaning of a polysemic term. For example, one basic term of the Tibetan grammatical tradition, *yi ge*, corresponds to the concept of a “phoneme, which can be expressed graphically,” and sometimes as a “syllable” or even “syllabographeme.” Thus, the concepts of phoneme, grapheme, syllable, and its components in the Tibetan tradition are not separated. A single concept denoted by the Tibetan term *yi ge* unites minimal units of linguistic sound (phonemes) with minimal units of the language graphic system (grapheme). In this and similar cases, the relation “to denote a concept” (and the inverse relation “to be denoted by a sign”) was used in the ontology. The

³² Grinev-Grinevich 2008: 25.

Tibetan term *yi ge* was modeled as a basic concept³³ and connected via this relation with the expression in Russian. Thus, we describe the meaning of the term *yi ge* as “a grapheme (syllabographeme); a linguistic sign denoting the phoneme.” Then, we connect this concept via the relation “to denote a concept” to the concept in Russian “any phoneme.” Parallel hierarchies are built for all types of Tibetan graphemes/phonemes. This allows us simultaneously reflect the dual meaning of the term and preserve the opportunity to participate in various semantic relations (for example, phonemes can be pronounced, graphemes can be written; graphemes can have graphic elements, but phonemes cannot; etc.).

6. Concluding Observations

Even in the initial stages of work, the development of the ontology for the Tibetan language demonstrated that ontologies are not language-independent, but should be individually developed for each particular language. The ontology of the Tibetan language reflects special features of Tibetan lexical, grammatical, and syntactic semantics, as well as the specifics of ordinary and special lexis functioning. Thus, the ontology is a formalized representation of the real-world knowledge expressed in the Tibetan lexicon and grammar.

Building an ontology of Tibetan allows the investigation of the structure of lexico-semantic fields and the meaning of Tibetan language elements, taking into account language facts from such areas as the structure of lexical systems, including polysemy and connotations, metaphorical compatibility, etc. This research will allow us not only to reveal features of the above-mentioned areas and to solve certain issues of system lexicography but also to understand the scope of differences that exists in this respect between classical and modern Tibetan.

Bibliography

Abdollahi, Mahdi *et al.* 2019. “An Ontology-based Two-Stage Approach to Medical Text Classification with Feature Selection by Particle Swarm Optimisation.” In *IEEE Congress on Evolutionary Computation (CEC)*, 119–126. Wellington: IEEE.

³³ Basic concepts are those with a large amount of semantic relations.

Dobrov, Alexey. 2014. "Semantic and Ontological Relations in AIIRE Natural Language Processor." In *Computational Models for Business and Engineering Domains*, edited by G. Seltak and K. Markov, 147–157. Rzeszow-Sofia: ITHEA.

———, Anastasia Dobrova, Pavel Grokhovskiy, Maria Smirnova, and Nikolay Soms. 2018. "Computer Ontology of Tibetan for Morphosyntactic Disambiguation." In *Digital Transformation and Global Society. DTGS 2018. Communications in Computer and Information Science* 859, 336–349. Accessed May 15, 2020. https://doi.org/10.1007/978-3-030-02846-6_27.

———, Anastasia Dobrova, Pavel Grokhovskiy, Nikolay Soms, and Viktor Zakharov. 2016. "Morphosyntactic Analyzer for the Tibetan Language: Aspects of Structural Ambiguity." In *International Conference on Text, Speech, and Dialogue*, edited by P. Sojka et al., 215–222. Accessed May 15, 2020. https://doi.org/10.1007/978-3-319-45510-5_25.

———, Anastasia Dobrova, Pavel Grokhovskiy, and Nikolay Soms. 2017. "Morphosyntactic Parser and Textual Corpora: Processing Uncommon Phenomena of Tibetan Language." In *Proceedings of the International Conference on Internet and Modern Society, IMS 2017*, edited by R. V. Bolgov, N. V. Borisov, L. V. Smorgunov, I. I. Tolstikova, and V. P. Zakharov, 143–153. Accessed May 15, 2020. <https://doi.org/10.1145/3143699.3143718>.

———, Anastasia Dobrova, Maria Smirnova, and Nikolay Soms. 2019. "Formal Grammatical and Ontological Modeling of Corpus Data on Tibetan Compounds." In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (2)*, edited by J. Diets, D. Aveiro, and J. Filipe, 137–143. DOI: 10.5220/0008162401440153.

Bolshoy entsiklopedicheskiy slovar 1998: *Bolshoy entsiklopedicheskiy slovar, Yazykoznanie* [Great Encyclopedical Dictionary, Linguistics], edited by Victoria N. Yartseva and Nina D. Arutyunova. Moscow: Nauchnoe izdatelstvo "Bolshaya Rossiyskaya entsiklopediya."

Grinev-Grinevich, Sergei V. 2008. *Terminovedeniye: Uchebnoe posobie dlya studentov vysshih uchebnykh zavedenij* [Terminology: Textbook for Students of Higher-Educational Institutions]. Moscow: Akademiya.

- Grokhovskii, Pavel, and Maria Smirnova. 2017. "Principles of Tibetan Compounds processing in Lexical Database." In *Proceedings of the International Conference on Internet and Modern Society, IMS 2017*, edited by R. V. Bolgov, N. V. Borisov, L. V. Smorgunov, I. I. Tolstikova, and V. P. Zakharov, 135–142. Accessed May 15, 2020. <https://doi.org/10.1145/3143699.3143718>.
- , Viktor Zakharov, Maria Smirnova, and Maria Khokhlova. 2015. "The Corpus of Tibetan Grammatical Works." *Automatic documentation and mathematical linguistics* 49 (5), 182–191. Accessed May 15, 2020. <https://doi.org/10.3103/S0005105515050064>.
- Gruber, Thomas R. 1993. "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition* 5 (2), 199–220.
- Jensen, Per Anker, and Jørgen Fischer Nilsson. 2006. "Ontology-based Semantics for Preposition." In *Syntax and Semantics of Prepositions*, edited by Patrick Saint-Dizier, 229–244. Dordrecht: Springer.
- Kang, Sin-Jae, and Jong-Hyeok Lee. 2001. "Ontology-based Word Sense Disambiguation by Using Semi-automatically Constructed Ontology." In *Proceedings of MT Summit VIII*, edited by Bente Maegaard, 199–220. Accessed May 15, 2020. <http://mtarchive.info/MTS-2001-TOC.htm>.
- Kasevich, Viktor B. 1996. *Buddizm. Kartina mira. Yazyk* [Buddhism. Picture of the World. Language]. Saint-Petersburg: Tsentr Peterburgskoye Vostokovedeniye.
- "Leipzig Glossing Rules." Max Planck Institute for Evolutionary Anthropology Department of Linguistics. Accessed May 20, 2020. <https://www.eva.mpg.de/lingua/resources/glossing-rules.php>.
- Lytvyn, Vasyl *et al.* 2017. "Classification Methods of Text Documents Using Ontology Based Approach." In *Advances in Intelligent Systems and Computing*, edited by Natalya Shakhovska, 229–240. Cham: Springer.
- Maslov, Yuriy S. 1998. "Glagol (Verb)." In *Bolshoy entsiklopedicheskiy slovar, Yazykovedeniye* [Great Encyclopaedical Dictionary, Linguistics], edited by Victoria N. Yartseva and Nina D. Arutyunova, 104–105. Moscow: Nauchnoe izdatelstvo "Bolshaya Rossiyskaya entsiklopediya."

- Sánchez-Cisneros, Daniel, and Gali F. Aparicio. 2013. "UEM-UC3M: An Ontology-based Named Entity Recognition System for Biomedical Texts." In *SemEval Vol.2*, edited by Suresh Manandhar and Deniz Yuret, 622–627. Atlanta: Association for Computational Linguistics.
- Sánchez-Pi, Nayat, Luis Martí, and Ana Cristina Bicharra Garcia. 2016. "Improving Ontology-based Text Classification: An Occupational Health and Security Application." *Journal of Applied Logic* 17, 48–58.
- Zhang, Yisun. 1985. *Bod rgya tshig mdzod chen mo* [Great Tibetan-Chinese dictionary]. Beijing: Minzu chubanshe.
- Zhou, Peng, and Nora El-Gohary. 2015. "Ontology-based Multilabel Text Classification of Construction Regulatory Documents." *Journal of Computing in Civil Engineering* 30 (4). Accessed May 15, 2020. [https://doi.org/10.1061/\(ASCE\)CP.1943-5487.0000530](https://doi.org/10.1061/(ASCE)CP.1943-5487.0000530).

