sciendo
DOI 10.2478/jazcas-2021-0030

FROM GRAPHEMATICS TO PHRASAL, SENTENTIAL, AND TEXTUAL SEMANTICS THROUGH MORPHOSYNTAX BY MEANS OF CORPUS-DRIVEN GRAMMAR AND ONTOLOGY: A CASE STUDY ON ONE TIBETAN TEXT

ALEKSEI DOBROV[1] – MARIA SMIRNOVA[2]
[1] LLC "AIIRE", Saint Petersburg, Russia
[2] Saint Petersburg State University, Saint Petersburg, Russia

DOBROV, Aleksei – SMIRNOVA, Maria: From graphematics to phrasal, sentential, and textual semantics through morphosyntax by means of corpus-driven grammar and ontology: A case study on one Tibetan text. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 319 – 329.

Abstract: This article presents the current results of an ongoing study of the possibilities of fine-tuning automatic morphosyntactic and semantic annotation by means of improving the underlying formal grammar and ontology on the example of one Tibetan text. The ultimate purpose of work at this stage was to improve linguistic software developed for natural-language processing and understanding in order to achieve complete annotation of a specific text and such state of the formal model, in which all linguistic phenomena observed in the text would be explained. This purpose includes the following tasks: analysis of error cases in annotation of the text from the corpus; eliminating these errors in automatic annotation; development of formal grammar and updating of dictionaries. Along with the morpho-syntactic analysis, the current approach involves simultaneous semantic analysis as well. The article describes semantic annotation of the corpus, required by grammar revision and development, which was made with the use of computer ontology. The work is carried out with one of the corpus texts – a grammatical poetic treatise *Sum-cu-pa* (VII c.).

Keywords: Tibetan language, computer ontology, Tibetan corpus, natural language processing, corpus linguistics, parsing

1 INTRODUCTION

This article discusses the development of a formal model (a grammar and a linguistic ontology) of the Tibetan language, including morphosyntax, syntax of phrases and super-phrasal units, and semantics that can perform the morpho-syntactic, syntactic, and semantic analysis. The engine is based on a consistent formal model of Tibetan vocabulary, grammar, and ontology, verified by and developed on the basis of a representative and manually tested corpus of texts, which includes the Basic Corpus of the Tibetan Classical Language [1] and the Corpus of Indigenous Tibetan Grammar Treatises [2], comprising 34,000 and 48,000 tokens, respectively [3]. Among the texts of our corpus, there are both prose and poetic texts.

Tibetan can reasonably be considered as one of the less-resourced languages. Despite the fact that scholars in different countries (Germany, United Kingdom, China, USA, Japan) are working on the tools for processing Tibetan texts, there is still no conventional standard for annotating a corpus of Tibetan language material. A number of recent studies were primarily aimed at developing solutions for such stages of Tibetan NLP as word segmentation and part-of-speech tagging. Some researchers use corpus methods to solve specific applied problems, as well as tasks in the field of history, literature, linguistics, and anthropology (e.g., [4], [5], [6]). Syntactic and semantic research of Tibetan has been comparatively weak.

The common problem of formal grammar development for less resourced languages is that in order to create an adequate formal model, a representative corpus of texts with reliable annotation must be created first, but in order to annotate a corpus, a formal model must already exist to form the basis of annotation. Thus, when working with the Tibetan language, we decided to implement an approach that allows us to gradually improve the formal grammar and develop the existing corpus. Initially, work was carried out simultaneously on all texts of the corpus. Various types of errors indicating typical problems (morphosyntactic ambiguity or lack of syntactic or semantic annotation) were analyzed and resolved taking into account all cases in the corpus representing a particular problem. When the problems became specific, we decided to analyze separate texts, sequentially improving linguistic software. The article describes the methods of working with the first text of the corpus – the Tibetan grammatical treatise *Sum-cu-pa* (VII c.) – the corrections required and the problems encountered. The choice of a poetic text is explained only by the fact that it is the shortest text in the corpus, consisting of 1356 tokens. It is convenient to use it to demonstrate the results of our work.

2 THE SOFTWARE TOOLS

This study is carried out within the framework of the AIIRE project [7] and with use of the technologies and tools of this project. AIIRE is a free open-source natural language understanding system, which is developed and distributed under the terms of GNU General Public License. This system implements the full-scale procedure of natural language understanding, from graphematics, through morphological annotation and syntactic parsing, up to semantic analysis.

2.1 Tokenization and morphological annotation

The module developed for the Tibetan language is designed taking into account the fact that since there are no separators between words in Tibetan writing, while morphology and syntax are significantly intermixed, the minimal (atomic) units of modeling (so-called atoms) are morphemes and their allomorphs, not words and their forms. Input string segmentation into such units (tokenization) cannot be

performed with standard tokenization algorithms, and is therefore performed in AIIRE by means of the Aho-Corasick algorithm (developed by Alfred V. Aho and Margaret J. Corasick [8]). This algorithm allows one to find all possible substrings of the input string according to a given dictionary. The algorithm builds a tree, describing a finite state machine with terminal nodes corresponding to completed character strings of elements (in this case, morphemes) from the input dictionary.

The Tibetan language module contains a dictionary of morphemes with their allomorphs, so that this tree can be created in advance at the build stage of the module and loaded as a shared library in the runtime, which minimizes its initialization time. The dictionary of morphemes contains grammatical and morphological attributes (grammemes) for each allomorph; these attributes are mapped onto classes of immediate constituents, so that the tree for the Aho-Corrasick algorithm contains just class and morpheme identifiers for each allomorph and doesn't need to store individual attributes. The module also contains a set of definitions that determines possible types of atoms (atomic units), possible attributes for each type of atom, possible values of each attribute, and restrictions on each attribute value.

Thus, AIIRE first builds all possible hypotheses of recognizing Tibetan atomic units in input strings, including overlapping substrings for separate hypotheses, and then brings them together immediately after they arise into trees of immediate constituents in all possible ways in accordance with the formal grammar, which models the Tibetan morphosyntax.

2.2 Syntactic parsing

The grammar is a combined grammar of immediate constituents and syntactic dependencies, which consists of the so-called classes of immediate constituents (CICs hereinafter). CICs are developed as python-language classes, with enabled built-in inheritance mechanism, and specify the following attributes: a semantic graph template which represents how the meaning of a constituent should be calculated from the meanings of its child constituents; lists of possible head and subordinate constituent classes; a dictionary of possible linear orders of the subordinate constituent in relation to the head and the meanings of each order; the possibility of head or subordinate constituent ellipsis; the possibility of non-idiomatic semantic interpretation [9, p. 146]. Currently, the formal grammar includes 507 CICs.

The formal grammar is developed in direct accordance with semantics, in a way that the meanings of syntactic and morphosyntactic constituents can be correctly calculated from the meanings of their child constituents in accordance with the Compositionality principle.

The Tibetan language module is integrated into AIIRE natural language processor, and the corpus texts are passed on for processing in unannotated form.

The results of linguistic processing are presented in the form of immediate constituent structures with semantic graphs, these structures forming the syntactic and semantic annotation of the corpus: the results of automatic text processing are loaded into the AIIRE corpus manager as the annotation of the corpus, upon which the corpus manager automatically searches for typical errors, indicating locations of incomplete annotation and possible inaccuracy. The four types of errors are: unrecognized units, combinatorial explosions, breaks in syntactic trees, and overlaps. Unrecognized fragments are those for which there are no syntactic trees in the annotation. Combinatorial explosions are cases of exponential growth in the number of parsing versions with respect to the length of the parsed text and, thus, the amount of its parsed ambiguous fragments increases. Breaks are positions in which the tree cannot be bound with any of its neighbours. Overlaps are fragments of text in which the syntactic trees overlap, not completely covering the text: a fragment covered by one tree includes the position of the beginning of the fragment covered by the next tree, but not the position at its end [10, p. 145].

This toolkit allows simultaneous work on the corpus annotation and on the improvement of the formal model behind this annotation, which is a new approach to the development of modules of the linguistic processor, ensuring continued verifiability of the formal model and its correspondence with the corpus material.

2.3 Semantic analysis

The ontology used for this research is a united, consistent classification of concepts that unite the meanings of linguistic units of the corpus texts, including morphemes and idiomatic morphemic complexes. To model a new concept, a researcher needs to create an expression entry in the ontology, and provide it with the meaning (translation) and description (or interpretation) in Russian.[1] The main source for establishing the basic meaning of each expression is a text or texts of the corpus where the expression is used. Regular use of the Tibetan explanatory dictionary helps to verify the choice of a Tibetologist, who edits the ontology. In some cases, translating and interpreting linguistic units (especially special terms) requires thematic dictionaries, thesauri and catalogues. In controversial cases, native speakers are involved.

The concepts are interconnected with different semantic relations. In addition to such semantic relations as synonymy, hyponymy, and hypernymy, the ontology models strictly specified relations between concepts such as the relation between a physical object and its parts (meronymy); between the agent and the actions that the agent can perform; between an action and objects towards which this action can

[1] 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.

be directed, etc. These relations allow semantic analysis of texts and lexical and syntactic disambiguation to be performed. The basic ontological editor is described with examples from the Tibetan ontology ([11], [12], [13]). As far as the authors of this article are currently aware, at the moment, AIIRE is the only system that actually implements not only word-sense, but also syntactic disambiguation by means of linguistic ontology without use of any statistical heuristics.

The ultimate goal of our project is to create a complete semantic annotation of all texts in the corpus. At the moment, 5230 concepts are modelled in the ontology, including the meanings of all lexical units (verbs, compounds and idiomatic expressions; 160 concepts in total) of the *Sum-cu-pa* grammar.

3 ANNOTATION DEVELOPMENT

The Thirty Verses (Tib. *Sum-cu-pa,* presumably 7th–9th centuries AD) is one of the first two Tibetan grammatical treatises that laid the basis for traditional Tibetan linguistics (Tib. *sgra'i rig-pa*). Tibetan proto-scientific texts have special structural features and methods of description, and use a large number of grammatical terms and special lexis. The characteristics of Tibetan poetic texts (omission of grammatical markers, ellipsis, adding syllables to comply with the poetic meter) also cause a number of difficulties in syntactic and semantic analysis and require updating of dictionaries and formal grammar development, along with the use of computer ontology.

3.1 New classes of atoms

Since the grammatical description in the *Sum-cu-pa* treatise begins with the structure of Tibetan syllables, after which the author describes the formation and meaning of various grammatical markers, it became necessary to create two new separate classes of atoms for letters and exponents of Tibetan morphemes and function words (e.g., the allomorph *gyi* of the morpheme *KYi* that expresses the genitive case meaning). The class Letter contains the letters of the Tibetan alphabet, and the class Exponent contains exponents of morphemes, which were used with metalinguistic meaning in the *Sum-cu-pa* grammar.

Letters and exponents act like nouns in the text – they can attach attributes (even to each other like in (1)); act as a subject or direct object of certain verbs. Separate classes of immediate constituents – entity argument (EntityArg[2]) and entity right argument (EntityRightArg) – were created for combinations "Letter/Exponent + intersyllabic delimiter" and "intersyllabic delimiter + Letter/Exponent". These classes, in turn, were embedded as arguments into transitive verbal phrases (TransitiveVP) and noun phrases with genitive (NPGen) respectively, which ensured correct syntactic parsing of sentences like (1).

[2] The names of the CICs from the formal grammar as they appear in syntactic graphs.

(1) སུ་ཡི་ཨུ་ཕྱིས་ནས། །དེ་ལ་གསུམ་པའི་དང་པོ་སྦྱར།
su yi u phyis nas / / de la gsu- pa ‘i dang-po sbyar
su GEN u remove-EL PDem DAT third GEN first join
‘After removing *u* [from the grammatical marker] *su*, add the first [letter] of the third [alphabet row] to it.’

All atoms that belong to the class Letter were assigned to the ontology concept *yi-ge* ‘grapheme’; while atoms that belong to the class Exponent inherit the concept ‘linguistic unit’. Thus, in order to avoid breaks in the annotation and ensure the correct semantic analysis of the genitive noun phrase from example (1) *su yi u* ‘*u* of [the marker] *su*’, the concept ‘linguistic unit’ was connected via the relation ‘to have a grapheme’, that is a subclass of the general genitive relation ‘to have any object or process’, with the concept *yi-ge* ‘grapheme’. The same concepts were used to limit the verb valencies.

3.2 Topicalized noun phrases

The *Sum-cu-pa* grammar is composed in heptasyllabic verses, united in *shlokas* (Sans. śloka, Tib. *sho-k+la*).[3] It is written in the most common meter that was used as a standard translation of Sanskrit *shlokas* and in much of the native poetry in classical Tibetan. The meter implies that every line should have seven syllables [14, p. 410]. In order not to violate the meter, the author of the treatise sometimes excessively uses the topicalizer *ni*. The text contains structurally identical phrases, in one of which there is a topicalizer, while in the other it is absent.

The topicalizer is used after the ordinary noun phrase in only nine out of twenty-four cases. In other cases, it is added excessively (also for filling out the meter) after function words denoting case meanings. For example, in (2) it is used after a noun phrase in the ablative.

(2) བདུན་པ་ལས་ནི་ཤ་མ་གཏོགས། །རྗེས་འཇུག་ཡི་གེ་བཅུ་རུ་འདོད། །
bdun pa las ni sha ma gtogs/ /rjes-’jug yi-ge bcu ru ‘dod/ /
seventh ABL TOP *sha* NEG-belong final_consonant phoneme ten TERM accept
‘As for [phonemes that are] from the seventh [row of the alphabet], all except the first letter belong to the final phonemes.’

For such cases the following CICs were created: topicalized noun phrase in ablative (TopicalizedAblativeNP); genitive (TopicalizedGenitiveNP); terminative (TopicalizedTerminativeNP); dative (TopicalizedDativeNP) and ergative (TopicalizedErgativeNP). In the ontology, the relation ‘to concern an object or process’ was created for topicalized noun phrases. This relation, in turn, was

[3] Element of a poetic text (analogue of a stanza). In the *Sum-cu-pa* treatise *shlokas* include from two to five seven-syllable lines.

connected via the relation ‘to have an object’ with the class of possible topics (that is, with any concept).

3.3 Zero nominalization

The term *zero nominalization* was suggested by N. Hill for morphologically finite forms occurring in syntactically nominal contexts [15, 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 [14, p. 305]. Several examples of this phenomenon can be found in the *Sum-cu-pa* treatise. In most of them, the right context indicates that a verb functions as a noun. Usually, the nominalizer *–pa* occurs only after the last of several verbs, while the nominalization of the preceding ones is guaranteed by the choice of the conjunction particle *dang* like in (3), since *dang* occurs only after nouns or noun phrases ([14, p. 241], [15, p. 5]).

(3) 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 (3), 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. The CIC poetic verbal noun (PoeticVN) was created for such cases in the formal grammar. This class was embedded in the CIC for noun phrases in the plural (InstanceNPPlural) and homogeneous noun phrases (InstanceNPGroup).

In examples (4) and (5), noun coordinators are used only once at the end of the passage.

(4) 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.’

(5) 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 (4), 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 this not a typical grammatical phenomenon for the Tibetan language, the special class PoeticHomogenVP was created and embedded into classes for verbal nominalization.

In example (5), 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. Disambiguation in this case will be discussed below (see section 3.5).

3.4 Equative verb omission

The equative verb *yin* expresses equation or identification of two patient participants (nouns or noun phrases of different length and complexity) [14, p. 255]. The *Sum-cu-pa* treatise demonstrates several omissions of the equative verb *yin* before the statement final particle *–O*. In most cases, the verb is omitted in a compound nominal predicate consisting of numeral like in (6).

(6) ཀཱ་ལི་སུམ་ཅུ་ཐམ་པའོ
kA-li sum-cu tham-pa ‘o
consonant thirty even FIN
‘Consonants [are in the amount] thirty even.’

In the formal grammar, there was already a class for the copula group, consisting of an equative verb and a noun phrase. For cases like (6), the CIC for copula group with elliptic verb (EllCopulaGroup) was created, in which quantitative noun phrases were embedded.

3.5 Compounds with complex structure

Modeling of Tibetan compounds’ meanings was the first ontological task, since most combinatorial explosions contained compounds. As a result of this work, a classification of Tibetan verbal and nominal compounds was created. These types and their formal grammatical and ontological modeling are described in detail in [9].

In some cases, one of the components of a compound is itself a compound. In the *Sum-cu-pa* treatise, even more complex structures were discovered. In example (7), we found five verbal roots following each other without any markers between them. Relying on the context and several most authoritative commentaries on the *Sum-cu-pa* grammar, this passage can be read in the following way:

(7) སྡེབ་པར་སྦྱོར་བ་ལེགས་པར་མཛད་པའི་མཁས་པ།
sdeb [pa r] sbyor-[ba] legs-[pa r] mdzad-[pa ‘i] mkhas [pa]

composite-NMLZ TERM join-NMLZ be_good-NMLZ TERM do-NMLZ GEN be_skilled-NMLZ
‘[those who are] skilled in good making joining [of words] for composition’

Even if we do not take into account zero nominalization of the last verbal root *mkhas* ‘to be skilled’, other verbs in this passage are obviously in subordinate syntactic relations of different types with the omission of various grammatical markers. Omission of grammatical markers may be considered acceptable in a poetic text. However, changing the whole formal grammar to ensure correct syntactic analysis of this passage will inevitably cause combinatorial explosions. In this regard, it was decided to model the whole passage as a compound.

According to the created model of Tibetan compounds, reconstructed syntactic relations allow to consider *sdeb-sbyor* and *legs-mdzad* as compound atomic verbal phrases with circumstance (CompoundAtomicVPWithCirc). 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 one atom (CompoundAtomicTerminativeNP) and the intersyllabic delimiter (the terminative case marker is omitted as is usual 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 *legs-mdzad*, this relation was established on the basic class of its nominal component *legs-pa* ‘being good’ – ‘any process’. Syntactic relations between *sdeb-sbyor* ‘poetry’ and *legs-mdzad* ‘to do well’ are the same as in compound transitive verb phrases (CompoundTransitiveVP), where the first nominal component is a direct object of the second verbal component. In turn, the syntactic relations between *sdeb-sbyor-legs-mdzad* and *mkhas-pa* are the same as those between the components of a noun phrase with genitive compounds (NPGenCompound). These cases of compounds with complex structure are not common, so it was decided not to change immediate constituents of the CICs CompoundTransitiveVP and NPGenCompound, but to create separate classes for compound’s groups.

3.6 Annotation error statistics

As mentioned above, the AIIRE corpus manager automatically searches and counts cases of unrecognized fragments, gaps between syntactic structures or their overlaps, and combinatorial explosions. Regular processing of the *Sum-cu-pa* grammar gives us the following statistics of these annotation errors (table 1). The third column also takes into account special cases of described changes, as well as some minor changes, which were not described here.

	1	2	3	4
	Before introducing the ontology	Before the improvements proposed	After the improvements proposed	Current amount
Amount of gaps (tokens)	151	18	7	9
Unrecognized (tokens)	10	0	0	0
Overlaps (tokens)	9	7	0	4
Combinatorial explosions (tokens)	0	0	0	0
Amount of gaps (sentences)	744	323	196	196
Unrecognized (sentences)	26	1	0	0
Overlaps (sentences)	96	59	26	31
Combinatorial explosions (sentences)	7	0	0	1

Tab. 1. The *Sum-cu-pa* grammar processing statistics

The statistics in Tab. 1 takes into account only text processing cases where the syntactic and the semantic analysis were performed simultaneously. Here, we do not provide statistics of annotation errors only for the syntactic mode parsing, because the processing of texts without semantic restrictions at the previous stages of work showed critical ambiguity at the level of syntax.

4 CONCLUSIONS AND FURTHER WORK

The fine-tuning of the automatic morphosyntactic and semantic annotation of the *Sum-cu-pa* grammatical treatise eliminates all unrecognized fragments in the text, almost completely eliminates combinatorial explosions, and significantly reduces gaps in annotation. The remaining breaks are caused by the lack of full semantic annotation of the text (in the syntactic parsing mode the number of gaps is much lower, but the number of versions of parsing becomes unacceptably large).

At the moment, most of linguistic phenomena observed in the text are explained in the current version of formal grammar. We take into account that some of them can be characteristic only for poetic texts or only for texts of a certain period of the Tibetan language development. Further work will mainly include development of semantic annotation of the Sum-cu-pa and completion of work with all the texts of the corpus (of different periods, poetic and prose).

ACKNOWLEDGEMENTS

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

[1] The Basic Corpus of the Tibetan Classical Language. (2019).

[2] The Corpus of Indigenous Tibetan Grammar Treatises. (2019).

[3] The corpus of Tibetan grammatical works. In Automatic documentation and mathematical linguistics, vol. 49, no. 5, pages 182–191. https://doi.org/10.3103/S0005105515050064.

[4] Wagner, A., and Zeisler, B. (2004). A syntactically annotated corpus of Tibetan. In Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon, pages 1141–1144.

[5] Semantic Roles, Case Relations, and Cross-Clausal Reference in Tibetan. URL: http://www.sfb441.unituebingen.de/b11/ b11corpora.html# clarkTrees.

[6] Grokhovskiy, P., Khokhlova, M., Smirnova, M., and Zakharov V. (2015). Tibetan Linguistic Terminology on the Base of the Tibetan Traditional Grammar Treatises Corpus. In P. Král and V. Matoušek (eds.), Text, Speech, and Dialogue. TSD 2015. Lecture Notes in Computer Science, vol 9302. Springer, Cham.

[7] Dobrov, A., Dobrova, A., Grokhovskiy, P., Soms, N., and Zakharov V. (2016). Morphosyntactic analyser for the Tibetan language: aspects of structural ambiguity. In International Conference on Text, Speech, and Dialogue, pages 215–222. DOI 10.1007/978-3-319-45510-5_25.

[8] Aho, A. V., and Corasick, M. J. (1975). Efficient string matching: An aid to bibliographic search. Communications of the ACM. 18, 6, pages 333–340.

[9] Dobrov, A., Dobrova, A., Smirnova, M., and Soms, N. (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, vol. 2: KEOD, Vienna, Austria.

[10] Dobrov, A., Dobrova, A., Grokhovskiy, P., Soms, N. (2017). Morphosyntactic Parser and Textual Corpora: Processing Uncommon Phenomena of Tibetan Language. In Proceedings of the International Conference IMS-2017, pages 143–153. DOI 10.1145/3143699.3143719.

[11] Dobrov, A., Dobrova, A., Grokhovskiy, P., Smirnova, M., and Soms, N. (2018). Computer ontology of Tibetan for morphosyntactic disambiguation. In D. A. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Y. Kabanov and O. Koltsova (eds.), Digital Transformation and Global Society, pages 336–349, Cham. Springer International Publishing. https://doi.org/10. 1007/978-3-030-02846-6_27.

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

[13] 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.

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

[15] Hill, Nathan W. (2019). Tibetan zero nominalization. Revue d'Etudes Tibétaines, no. 48, Paris.