Tibetan Linguistic Terminology on the Base of the Tibetan Traditional Grammar Treatises Corpus

Pavel Grokhovskiy(✉), Maria Khokhlova,
Maria Smirnova, and Victor Zakharov

Saint-Petersburg State University, Saint-Petersburg, Russia
{plgr,2321781}@mail.ru, khokhlova.marie@gmail.com, vz1311@yandex.ru

Abstract. The paper is devoted to Tibetan grammatical terminology. For this purpose Tibetan grammatical works corpus was created. At the same time Russian translations of the works were added to the corpus, so it is factually a parallel Tibetan-Russian corpus. The corpus represents the collection of grammar treatises of the Tibetan grammatical tradition formed in VII-VIII c. The corpus is useful to researchers of the Tibetan linguistic tradition as well as to those specialized in linguistic studies of classical and modern Tibetan and its teaching. On the basis of corpus a specific grammatical lexical database is created. The database will be useful both to tibetologists and general linguistics specialists.

Keywords: Linguistical terminology · Tibetan language · Corpus linguistics · Parallel corpus · Morphology · Tagging · Lexical database

1 Introduction

The project focuses on the creation of the Tibetan grammatical terminology database and Tibetan traditional grammar treatises corpus. The origin of linguistic tradition in Tibet is dated back to the creation of the first grammatical treatises "Sum cu pa" and "Rtags kyi 'jug pa" (VII-VIII c.).

The Tibetan linguistics is mainly based on grammars created by Buddhist scholars and is thus highly connected with Indian tradition. Methods of language characterization and analysis are greatly different from those of Western linguistics. Modern Tibetan scholars continue to follow and develop the Tibetan grammatical tradition. The corpus includes the basic grammar treatises and commentaries, which are considered to be the most important grammatical works within the Tibetan grammatical tradition. On the basis of the corpus a specific grammatical lexical database is created which will be useful both to tibetologists and general linguistics specialists.

P. Král and V. Matoušek (Eds.): TSD 2015, LNAI 9302, pp. 326–333, 2015.
DOI: 10.1007/978-3-319-24033-6_34

2 Modern Corpus Linguistics of the Tibetan Language

Despite the fact that scholars in different countries (Germany, Great Britain, People's Republic of China, USA and Japan) are engaged in working out of Tibetan texts corpus presentation, still there is no common standard for it.

Last four conferences of the International Association for Tibetan Studies (IATS Seminar) included section «Tibetan Information Technology», where computer technology projects in the field of Tibetan studies were represented. They also include projects focusing on the creation of Tibetan corpus.

The creation of Tibetan language corpora abroad has just begun. The cooperative research project 441 under the guidance of B. Zeissler in Eberhard Karls University (Tübingen, Germany) included the subproject B11 «Semantic roles, case relations, and cross-clausal reference in Tibetan» (2002-2008) [1]. In 2012 U. Pagel from Department of the Study of Religions and N. Hill from Department of China and Inner Asia and Departments of Linguistics began development of the Tibetan corpus to contain 1 million syllables. Its texts cover three historical periods of the Tibetan language: preclassical, classical and modern[1].

The first difference between the corpus of the Tibetan traditional grammar treatises and the projects mentioned above is the development of special system of linguistic tags [2][3]. The second difference is related to the involved materials. All texts represent one of the traditional Tibetan sciences – linguistics.

3 Corpus Structure

The project has two main tasks: creation of a parallel corpus of the Tibetan grammatical treatises with Russian translation and creation of a specific grammatical lexical database with frequency characteristics and semantic relations.

Tibetan texts and Russian translations in the corpus are aligned by sentence breaks of the Tibetan part. Words of the Tibetan part are tagged morphologically. It should be noted that segmentation of Tibetan texts is a sophisticated problem because according to the traditional Tibetan orthography only syllable borders are marked (Tibetan syllable coincides with a word form approximately in 95 cases out of 100).

There are special programs for automatic segmentation, for example, Hunalign, Vanilla, etc. However, the process of their implementation for the needs of the Tibetan language is quite difficult and not included in the project tasks. Comparatively small corpus volume and the necessity of manual tokenisation argue for manual segmentation and tagging.

4 Corpus Annotation

Tibetan text undergoes morphological tagging (lemmatization, part-of-speech tagging, grammatical annotation of verb forms, eliminating of grammatical

Table 1. Tags for function words in the Tibetan language

No.	Tag	Function word	Example
1	Cj	Conjunction	dang
2	Pp	Postposition	drung du
3	Erg	Ergative	allomorphs kyis, gyis, gis, s, yis
4	Com	Comitative	dang
5	Dat	Dative	la
6	Loc	Locative	na
7	Dest	Destinative	allomorphs tu, du, ra, ru, su
8	Abl	Ablative	las
9	El	Elative	nas
10	Comp	Comparative	allomorphs pas, bas
11	Gen	Genitive	allomorphs kyi, gyi, gi, ‘i, yi
12	Fin	Final particle	allomorphs go, ngo, do, no, bo, mo, ‘o, ro, lo, so, to
13	Top	Topicalizing particle	ni
14	Ind	Indefinite particle	allomorphs cig, zhig, shig
15	Emph	Emphatic particle	allomorphs kyang, yang, ’ang
16	Quant	Quantifier (’that much’, etc.)	tsam, kho na, ’ba’ zhig, snyed
17	Pl	Plural marker	rnams
18	Quot	Quotation marker	allomorphs ces, zhes

homonymy). The Tibetan language tag system is developed. List of tags for the most widely used function words is given in Table 1.
Corpus texts are also provided with metadata about genre, date of creation, author.

Every word is represented by the following data: word form in Tibetan script, word form in Latin transliteration, lemma in Tibetan script, lemma in Latin transliteration, part-of-speech tag, terminological tag (Table 2). The program tool for automatic tagging TreeTagger is supposed to be adapted to the Tibetan language. In this regard the manually tagged corpus will be used as training corpus.

5 Corpus Search and Usage

The system Sketch Engine (as well as Nosketch Engine) is used as a corpus manager. The corpus can be accessed at http://corpora.spbu.ru. The corpus interface allows searching and filtration by all elements of annotation as well as brief and broad concordance and word lists creation. To create a frequency list all the above mentioned elements and their combinations could be used as attributes.

Currently the corpus contains 33913 tokens. Using corpus it is possible to search word forms, lemmas in the Tibetan script and Latin transliteration as well as collocations for given words and phrases, translations in parallel texts, operate with statistic data, etc. For an extended search the regular expression language is used.

[1] http://www.soas.ac.uk/news/newsitem73472.html as accessed on 29. 03. 2015.

Table 2. Fragment of aligned text

Word form (Tibetan sript)	**Word form** (transliteration)	**Lemma** (Tibetan script)	**Lemma** (transliteration)	**Part-of-speech tag**	**Terminological tag**
<s>					
<align>					
དེ་	de	དེ་	de	P	
ནི་	ni	ནི་	ni	Top	
སྡུད་	sdud	སྡུད་	sdud	VN	Gram L TGrMark
དང་	dang	དང་	dang	Cj	
འབྱེད་པ་	’byed pa	བྱེད་	byed	VN	Gram L TGrMark
དང	dang	དང	dang	Cj	
།།	//	།།	//	Punct	
རྒྱུ་མཚན་	rgyu mtshan	རྒྱུ་མཚན་	rgyu mtshan	N	Gram GenLex TGrMark
ཚེ་སྐབས་	tshe skabs	ཚེ་སྐབས་	tshe skabs	N	Gram L TGrMark
གདམས་ངག་	gdams ngag	གདམས་ངག་	gdams ngag	N	Gram GenLex TGrMark
ལྔ	lnga	ལྔ	lnga	Num	
འོ	’o	འོ	’o	Fin	
།།	//	།།	//	Punct	

The corpus of the Tibetan traditional grammar treatises could be used both for the purposes of theoretical and applied linguistics. The corpus is a source for lexicography, semantics and grammar investigations. Such corpus opportunities as statement comparison, search of lexical units equivalents allow to reduce time of working with teaching and information materials (e.g. dictionaries). Thus the corpus is a useful tool for translation and language teaching.

6 Lexical Database of the Tibetan Grammatical Treatises Corpus

6.1 Special Tagging of Grammatical Terminology

It is not typical for the Tibetan linguistics to emphasize such traditional subdisciplines of Western linguistics as phonology, morphology and syntax. Basic terms of the Tibetan grammatical tradition denote basic units of different language levels [4].

Most Tibetan authors begin their grammatical works with the description of the Tibetan alphabet, different types of graphemes and corresponding phonemes, rules of syllable composition and grapheme/phoneme combination as well as morphonological rules. Tibetan grammars also contain the description of function words and morphemes.

The Tibetan linguistic tradition borrowed Indian idea of seven cases. In Indian linguistics cases are connected with the kāraka category which represents an intermediate level between semantics and morphology. This system of kāraka categories was also borrowed by Tibetans.

6.2 Tags for Grammatical Terminology

Elements of traditional grammatical metadescription such as terminological categories (phonological, syntax terms), Sanskrit equivalents for loans, links to synonyms, hyperonyms, hyponyms are added to the lexical database as well as scientific commentaries.

The use of special grammatical tags given in Table 3 makes it possible to divide different terminological fields: grammatical terminology (tag Gram) and terms of traditional sciences (tag GenScien).

Certain tags stand for models of terms origin: by terminologisation of common words (tag GenLex) or through borrowing (tag L).

The tag TBas is used for basic grammatical terminology. Polysemy is the main feature of the Tibetan terminology in general and basic grammatical terms in particular. Therefore one of the main tasks was to separate phonological terms (tag TPhon) and terms for different types of graphemes (tag TGra).

Grammatical terms imported from the corpus through the use of special grammatical tags form the lexical database of Tibetan grammatical terminology, which contains additional information about the origin language for loans, foreign equivalents, way of borrowing (phonetic or semantic borrowing, calquing, hybrid terms), etc. There is more than 2000 occurrences of grammatical terms in the current corpus database.

Table 3. Tags for grammatical terminology

Characteristic of classification	**Tag**	**Meaning**
Terminological field	Gram	term of the Tibetan grammatical tradition
	GenScien	general scientific term
Origin	GenLex	term of Tibetan origin
	L	borrowed term
Type of terminology	TBas	basic grammatical term
	TPhon	phonological term
	TGra	grapheme type
	TGrMark	name of auxiliary morphemes and lexemes
	TCGr	case grammar term

6.3 Structure of Lexical Database

Lexical database of the Tibetan grammatical treatises corpus contains lexical units selected from the Tibetan part of the corpus by appropriate tags.

TEI recommendations (Text Encoding Initiative) are taken as a methodological basis for database exchange format [5]. It is important that TEI has tags for relation links to create network data representation in linear XML files.

Let's describe lexical database representation template in XML format according to database structure. This template has several divisions and representation levels.

Lexical unit level Lexical unit of the database in XML begins with record:

```
<entry n="1" type="lex">
<term>ཨཱ་ལི་</term>
<pron>A li</pron>
```

where index number of a lexical unit (n="1"), its type (type="lex" – lexical unit) and entry word ཨཱ་ལི་ in Tibetan script (tag <term>) and transcription (tag <pron>) are given.

It is followed by the block with grammatical information (tag <gramGrp>):

```
<gramGrp>
    <pos> N </pos>
</gramGrp>
```

which contains part-of-speech tag (tag <pos>) and additional grammatical information (tags <gen>, <flex>, etc.).

The same level includes one or several etymological information blocks (tag <etym>):

```
<etym n="1">
    <lbl> Sumcupa </lbl>
    <date> 8|$^{\rm th}$| c.</date>
</etym>
```

which contains sequence number of the etymology block, etymological tag (tag <etym>), source of information (tag <lbl>), date of the first usage (tag <date>). Also other tags, like <lang> (language, from which a word was presumably borrowed), could be used.

Link level and example level are represented in the same way. In the end data are converted from exchange format into Microsoft SQL on the Microsoft.NET platform.

6.4 User Interface

The database interface is a window application powered by Microsoft.NET and closely integrated with a system core. During the interface development the following tasks were set:

1. searching all lexemes in the database;
2. displaying lexemes in a convenient form;
3. manual adding of new lexemes;
4. editing of available lexemes;
5. loading (importing) lexical unit records from XML to TEI format;
6. saving (exporting) database records into TEI format.

Functions of the grammatical lexical database are as follows:

1. statistical data where appropriate;
2. to retrieve content of an entry for a given word;
3. to retrieve all synonyms for a given word;
4. to retrieve all hyponyms for a given word;
5. to retrieve all hyperonyms for a given word;
6. to show frequency in the database for a given terminological tag;
7. to show frequency in the corpus for a given terminological tag;
8. to retrieve all lexemes marked by a given terminological tag.

7 Future Works

In the future the corpus is supposed to be provided with syntactic annotation, extended and developed in a more extensive corpus of Tibetan texts including those dedicated to other traditional Tibetan sciences: Buddhist religious doctrine, logic, medicine, craft, poetics, synonymics, prosody, astrology and drama.

8 Conclusion

Pilot version of the Tibetan grammatical treatises corpus will be useful to all researchers of Tibetan grammatical written texts. Nowadays there are no available Tibetan language corpora aligned and translated into Russian. Therefore the corpus also could be useful for linguistic research, Tibetan language study and teaching.

Frequency dictionary of Tibetan lexical units (grammar terms) and semantic analysis of the lexical database will form a sort of linguistic ontology that includes hyponyms and hyperonyms, polysemic words and synonyms.

All this will allow to analyze the development of language in general and structure of terminological fields in particular, and to estimate the terminologisation degree of common words.

Acknowledgement. This corpus and database development was supported by the grant No. 13-06-00621 «The Pilot Version of Tibetan Grammar Texts' Electronic Corpus» by Russian Foundation for Basic Research. Authors also acknowledge Saint-Petersburg State University for a research grant 2.38.293.2014 «Modernizing the Tibetan Literary Tradition» which enabled the conceptual study of the original Tibetan texts.

References

1. Wagner, A., Zeisler, B.: A syntactically annotated corpus of Tibetan. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisboa, May 2004

2. Grokhovskiy, P.L.: Kategorii skazuemosti i nominalizatsii deystviya (substantivno-ad"ektivnye formy) v klassicheskom tibetskom yazyke. In: Guzev, V.G., (ed.) Ocherki po Teoreticheskoy Grammatike Vostochnykh Yazykov: Sushchestvitel'noe i Glagol, Izdatel'skiy dom SPbGU, pp. 269–288 (2001)
3. Grokhovskiy, P.L.: Grammatika imeni sushchestvitel'nogo v klassicheskom tibetskom yazyke. In: Guzev, V.G. (ed.) Ocherki po Teoreticheskoy Grammatike Vostochnykh Yazykov: Sushchestvitel'noe i Glagol, Izdatel'skiy Dom SPbGU, pp. 76–91 (2001)
4. Smirnova, M.O.: Bazovye terminy tibetskoy grammaticheskoy traditsii. In: Vestnik Sankt-Peterburgskogo universiteta. Seriya 13. Vostokovedenie. Afrikanistika. Vypusk 1, SPb, pp. 23–34 (2014)
5. Burnard, L., Bauman, S. (eds.) TEI P5: Guidelines for electronic text encoding and interchange (2010)