Construction and Application of Grammatical and Lexical Data for Sanskrit and Tibetan Parallel Texts

By Toru Aiba and Kyoji Oide
Graduate School of Information Sciences, Tohoku University.
Aramaki Aza Aoba 09, Aoba-Ku Sendai 980–8579, Japan.
{aiba,k-oide}@human.is.tohoku.ac.jp

Researchers of Indian Buddhism rely not only on Sanskrit but also translated Tibetan and Chinese texts. Tibetan translations are especially important being literal translations.

We are collecting two types of data, parallel texts in Sanskrit and Tibetan, are the most fundamental resources for researchers. We manually parsed them, split them into sentences and aligned the translations.

We also extract information from the original texts, such as parallel lexicons and grammars. As to lexica, we confirm about 60% of the words in the actual text may be handled in currently available lexicons. Our grammar resources are incomplete. We will have to collect further information concerning external sandhi and verb roots derivatives.

1. Introduction

Many of the original Sanskrit texts concerning Indian Buddhism have been lost. This has forced relevant researchers to rely on translated Tibetan and Chinese texts as substitutes. A large number of these texts, many translated over 1000 years ago, are extant.

For this reason Indian Buddhism researchers are forced to use both original Sanskrit texts and their translated texts in various languages. Suzuki[10)] tried to demonstrate the source of an additional portion of the chapter *TathAgatAyuH-pramANanirdezaparivarta*[1] in *SuvarNa-prabhAsa (Suv)* was a part of the Sanskrit text *MahAmeghasUtra (MSS)*. However, the original Sanskrit text of *MSS* is lost. In his comparative analysis of seven texts which included:

- translated Tibetan and Chinese texts of *MSS*.
- original Sanskrit text, translated Tibetan text, and two different Chinese texts of *Suv*.
- *JAtaka*, a Pali canon, which has a similar paragraph to the target portion.

he focused on the relationship between the assumed original Sanskrit text and its translations. First, he presupposed the additional portion of Sanskrit *Suv* was a prototype of the lost paragraph of *MSS*. Then he contrasted the translated texts of *MSS* with other texts, and reflected their differences into the supposed Sanskrit paragraph of *MSS*. Suzuki's research reveals the importance of the translated texts even when researchers can only refer to sentences of the original Sanskrit text.

Our purpose here is to construct a computer environment in which these languages can be handled seamlessly in terms of e-texts. Since this project is still in its early stages, we are now collecting relevant information that has already been

[1] In this article, we transliterate Sanskrit and Tibetan words into Roman alphabet. We follow the scheme of Harvard-Kyoto for Sanskrit and E-Wylie for Tibetan. They are listed in 1).

digitalized in order to confirm what linguistic information will be desirable.

We are aiming at two types of data. The first is information useful for researchers, such as the Sanskrit text of *Saddharma-puNDarIka (SDP)* and its translation in Tibetan. The second is information needed for computer processing, including the correspondence tables for verbal roots and stems in Sanskrit, and parallel lexica for Sanskrit and Tibetan.

2. Parallel Texts of SDP

2.1 Constructing Parallel Texts

Indian Buddhist studies are required to handle texts in many languages, including Sanskrit, Pali, Chinese, Tibetan, Mongolian, Japanese etc. Currently we are using texts in two languages, the Sanskrit and translated Tibetan texts. It is well known that the translated texts in Tibetan are stricter word for word translation than any other versions. We think this characteristic of Tibetan texts is a suitable selection for our research’s first object.

Sanskrit	**(22.2.15-22.2.16)**
sa ca sarvasattvapriyadarzano bodhisattvo mahAsattvas tasya bhagavataH pravacane duSkara-caryA abhiyukto abhUt/	
Tibetan	**(273b3-237b4)**
byang chub sems dpa’ = sems dpa’ chen po = sems can thams cad kyis mthong na dga’ ba = de ’ang/ bcom ldan ’das = de – ’i = gsung rab – la = dka’ ba byed pa – la = brtson pa – r = gyur te/	

Fig. 1 Data of eDic

We are participating in the “eDic” [9,11], which tries to construct a sentence-level contrastive dictionary of Tibetan and Sanskrit. This dictionary will offer users parallel sentences of the existing Sanskrit texts and its Tibetan translations. In other words, this project intends to provide a useful tool to researchers such as Suzuki[10] mentioned above, who study Indian Buddhist literature.

In this project, we are constructing parallel e-texts of *SDP* of Sanskrit and Tibetan. Our digitalized texts are based on the revised Sanskrit edition by Wogihara[14] and the Lhasa edition in Tibetan Tripitaka. Our data structure is shown in Figure 1. This example has two features. First, texts are split into sentences, and all the parallel sentences of Sanskrit and Tibetan are manually correlated. Second, each Tibetan sentence is cut into words with an ‘=’ mark, and then all postpositional particles are cut off from their content words by a ‘−’ mark. Since a digitalized dictionary of Classical Tibetan is not yet available, these symbols are necessary for a bilingual lexicon of Tibetan text.

Arranging parallel data, as shown in Figure 1 is very complicated and requires a high degree of competence in both languages. In addition to this hardship, it is hard to collect collaborators, since students of classical Sanskrit and Tibetan are few in number.

But in the near future, along with the accumulation of digitalized information concerning Sanskrit and Tibetan, we hope to be able to construct an automatic processing system, or at least tools to aid the splitting of texts into sentences, and the recognition of Tibetan word formation.

2.2 Applying Parallel Texts

We have created a trial search system which is based on parallel texts already prepared in order to demonstrate how the efficiency of the arrangement of our data.

The mechanism is very simple; it displays the paired sentences, which include all the search keywords from the user input in Tibetan.

Regarding the lettering, we use the transliteration in Roman alphabets for both Sanskrit and Tibetan, "Aiba's scheme" from 1). One of the reasons is that prepared devanagari or Tibetan letters are not yet in general use in the computer environment. Another reason is the existence of users who are not familiar with these native letters. We hope such users will also be able to understand or at least browse contents with the help of own system. However, users sometimes want to get search results in native letters such as devanagari and Tibetan. To comply with this request, we employ the Multilingual Imager for Web, "Mojiyaki"[7], which creates an image file from the various alphabets. When users require native letters, our system will produce HTML files which dynamically link to the "Mojiyaki" site from which they can get image files they need.

Throughout this process we have strongly felt that in order to seamlessly handle Sanskrit and Tibetan texts as well as to enable users to get the results they desire to a more complicated degree, we must collect an exhaustive parallel lexicon and grammar of both Sanskrit and Tibetan. In fact, users are now restricted to search our parallel texts via Tibetan only. They can not use Sanskrit for search keys yet, since our accumulated information is not sufficient for the more complicated word forms of Sanskrit. On the other hand, we wish to add to the system so that similar or parallel sentences in both Sanskrit and Tibetan, can be sought interchangeably. These processes require a great accumulation of many kinds of language information. Consequently, apart from collecting parallel texts, we have started to collect parallel lexicon and grammatical information.

3. Parallel Lexicon

3.1 Mvyut and TBTY

Several parallel lexicons of Sanskrit and Tibetan already exist, though digitalized ones for unrestricted use are few and far between. In this article, we use these two lexicons:

- *MahAvyutpatti (Mvyut)* [6]
- *Chinese-Sanskrit-Tibetan Table of Buddhist Terminology based on the YogacArabhUmi (TBTY)* [15]

3.1.1 Mvyut.

This is a digitalized version[6] based on Sakaki's work[8]. *Mvyut* was constructed by Imperial Order of Khri lde srong brtsan, a Tibetan king during the ninth century. The aim of this lexica is to prevent Buddhist terms from being confused when Sanskrit Buddhist texts are translated into Tibetan. Therefore its content is mainly Buddhist technical terminology. It contains about 9600 Sanskrit items and their Tibetan translations.

3.1.2 TBTY

This index[15] is the electronic version of the index constructed by Yokoyama and Hirosawa[16]. Its original base, the text *YogacArabhUmi* was compiled in the third- or fourth- century and is one of the most important canons of VijJAnavAdins of MahAyAna Buddhists. The electronic version contains about 26000 word pairs.

3.2 Adaptation of Mvyut

Mvyut itself and its revised text[8] both sometimes lack a strictness in description. This makes it difficult to rely on an unchanged version of the *Mvyut*, as the index *TBTY* might provide. So before using it, *Mvyut* must be adapted to our purpose. We have discover that difficulties are posed by both the *Mvyut* itself, and the construction of Sakaki's edition.

3.2.1 Problems of Mvyut itself

In Tibetan texts, alternative conjunctions "'am" etc., are included in the lexicon; creating confusion. As an example, "'di ltar ram rung" means either "'di ltar" or "rung". In most cases splitting these conjunctions automatically is enough. But these are some exceptions, such as "mi'am ci", a word meaning "an ugly people". We enumerate the possible divisions and the original string to avoid the mistake of division. We take it that for the time being, increasing the size of the underlying data will be more important than obtaining a higher degree of accuracy in the data to improve the accessibility of items in *Mvyut*.

In Sanskrit, most items are declined, usually in the Nominative case. Before employing our chosen lexicon, we have to identify the word stems. Because of the lack of digitalized information on Sanskrit, this identification itself may be only approximate.

3.2.2 Problems of Sakaki's edition

In Sakaki's edition, round brackets used in 984 items are confusing. In some cases they mean an item is omissible, such as "sgra sgrogs (kyi bu)", which means either "sgra sgrogs" or "sgra sgrogs kyi bu". In other cases it suggests a variant, for example, "don gyi (gyis) go ba" denotes either "don gyi go ba" or "don gyis go ba". We change the former round brackets in 237 items to square brackets to avoid this confusion. Latter brackets which suggest variants offer further problems since the scope of the variants are vague. In the case of "don gyi (gyis) go ba", it is difficult to decide of which syllable "gyis" is the variant without some knowledge of Tibetan. To cope with this problem, we assume "the number of syllables included in round brackets may correspond to the number of syllables of variant readings", and this assumption enables us to automatically define the scope of variants. There are some exceptions and they are being corrected manually.

3.3 Current Usability for Lexicon

3.3.1 Method

We tried to investigate the usability of our lexicons from the point of view of 1) how many words in text can be covered by lexicon; 2) how reliable the description of word pairs.

To evaluate our existing lexicons, we use a Sanskrit-Tibetan corpus in the twenty second chapter of *SDP*. The basic data structure is similar to that shown in Figure 1, though, in the Sanskrit part, we add the stem information for each word, i.e. the stem of "bodhisattvo" is "bodhisattva". First, for each word stem in the Sanskrit text of *SDP*, we confirm whether the word stem is described in our lexicon. If the word is found, we can get its Tibetan equivalents. Next, we search the paired Tibetan sentences to discover if one of the translated words would be found. When this verification process, Sanskrit to Tibetan is complete, we reverse the test, Tibetan to Sanskrit.

Table 1: Variety of Words in both lexicons

	Sanskrit	Tibetan
Mvyut	10116	13592
TBTY	17752	32298

Prior to our experiment, in Table 1 we show the variety of words found in both lexicons. We are not surprised to discover the number of Tibetan words is more than the Sanskrit in both lexicons as the two languages are so different, In the case of *Mvyut*, as mentioned in the former section, the Tibetan translation for one Sanskrit item sometimes consists of several items, which increases the number of Tibetan entries. On the other hand, a partial lack of Sanskrit's original source texts contributed to the smaller number of Sanskrit words in *TBTY*; a reliable edition for the whole text of *TBTY* is only published in Chinese and Tibetan translation. Sanskrit editions are partial.

Table 2: Coverage of Lexicon

		Mvyut	*TBTY*	*Merged*
Skt/Tib	ALL: 2400			
	OK	706 (29%)	1230 (51%)	1388 (57%)
	NG_{item}	1375 (57%)	541 (22%)	432 (18%)
	NG_{pair}	319 (13%)	629 (26%)	580 (24%)
Tib/Skt	ALL: 2215			
	OK	689 (31%)	1179 (53%)	1335 (60%)
	NG_{item}	864 (39%)	630 (28%)	434 (19%)
	NG_{pair}	662 (29%)	406 (18%)	446 (20%)

Table 3: Itemizations of NG_{item} in Sanskrit

Sanskrit (All:432)			
Noun(Name)	Pron	Verb	Others
267(130)	25	48	97

3.3.2 Results

The result is shown in Table 2. In this Table 2, *OK* the number of words which are found as an item in the lexicon, and whose companion in lexica is found in the parallel sentence. NG_{item} denotes the number of words which are not found in the lexicon. NG_{pair} is the number of words which are found in the lexicon, but the paired words are not found in their parallel sentence. When NG_{pair} is low, the descriptions of word pairs in the lexicon may be less reliable. Therefore NG_{pair} shows us the quality of the lexicon.

3.3.2.1 Analysis of the Merged

Table 3 itemizes NG_{item} in Table2. In Table 3, we focus on the result of the *Merged*.

Noun, *Pron*, *Verb* are the missing words, and they are respectively classified by their parts of speech. *Name* in *Noun* refers to proper nouns such as "sarvasattvapriya-darzana", a name of a Bodhisattva.

When searching, pronouns etc. are less important than any other elements. There are not many pronouns so it is not difficult to supplement them manually.

Others is mainly concerned with the problem of dividing words. For example, a word "atha" in Sanskrit is a particle used at the beginning of a paragraph. However, this "atha" often appears with "khalu", a particle implying "certainly", "indeed", etc., so some lexicons including Ejima[5)] classify "atha khalu" as a different item from "atha". In the case of Sanskrit, such differences in the policies of dividing a sentence into words become an important issue, since words and phrases often adhere to each other.

In contrast with Table 5, Verb occupies 10% of the all missing words. To discuss this matter, we shall focus on the two topics below. One is the complexity of the combination of verb-roots and preverbs. “anu-pra%dA” [2] , for example, is not included in the lexicon. But when we ignore the preverb “anu-pra” and pay attention only to the verb root “%dA”[3], we will find the verb root and its Tibetan companion “sbyin pa”. Unfortunately, such attempts are often unsuccessful, since preverbs are often used to change the meaning of verb roots. For example, a Tibetan translation of “A%dIp” might be “’bar ba”, but “%dIp” is paired with “ston pa” in *TBTY*. However, the cases in which the combination of preverbs prevent the discovery of existing verb roots were 27 (56% of the all missing verbs) in all. We will have to consider the way to deal with preverbs.

The second problem is related to the derivative words from verb roots. In Sanskrit, many nouns and adjectives derive from primitive verb roots. For example, a noun “anupradAna” is a derivative of a verb “anupra%dA”. In our case, “anupradAna” is included in the lexicon but “anupra%dA” is not. Such missing verb, though derivatives exist, amount to 24 (50%). 19 words revealed both problems.

This derivative strategy may be applicable to the missing noun. For example, a missing noun “vimocaka” and “vimocana” in the lexicon are linked because they both derive from the same verb “vi%muc”. However, using this idea may be risky, since their Tibetan pairs have different meanings[4].

3.3.2.2 Comparison between Mvyut and TBTY

In Sanskrit, the number of searchable word items (OK + NG_{pair}) via *TBTY* is 1859, consisting of 78% of the all words in the 22nd chapter of *SDP*. On the other hand, there were 1025 (43%) searchable words in *Mvyut*. The ratio of *TBTY* and *Mvyut* (1.81) is comparable to that of the variety of words in lexicon, as shown in Table 1. The rarity of searchable Sanskrit word items prevents us increasing the number of NG_{pair} of *Mvyut*.

Table 4: NG$_{pair}$ of Tib/Skt via Mvyut

Tib/Skt (All:226)		
Case (Pron)	Word (Deriv)	Verb
120 (97)	79 (22)	27

In Tibetan, it is notable that the number of searchable word items via *Mvyut* (1351) is very close to that of *TBTY* (1585), though the NG_{pair} of *Mvyut* become very high. A partial itemization of the Tibetan words whose pair is found in *TBTY* but not included in *Mvyut* is shown in Table 4. In this table, *Case* represents the number of failures to cope with declensions or sandhi of nouns and pronouns, among which *Pron* alone amounts to 97. This failure is more than the half of all the mistakes, since we have not yet provided grammatical information concerning pronouns. The remaining 23 examples of *Case* are the nouns whose declension pattern seem comparatively rare. Apart from *TBTY*, the characteristic problem of *Mvyut* lies in declension. This is the main reason why NG_{pair} of Tib/Skt in *Mvyut* becomes high. *Word* is the number of failures in which the lexical definition has no counterpart in in *SDP*. In *Word*, *Deriv* represents the words having related but not exactly corresponding partners. *Verb* are verbs.

[2] In this article, “%” in Sanskrit string means the beginning of verb root.

[3] When we ignore only a part of the preverb such as “anu”, we still find “pra%dA” and its Tibetan translation.

[4] The Tibetan for “vimocaka” is “thar bar byed pa”, and “vimocana” is “rnam par dgrol ba”.

Mvyut doesn't contain Sanskrit verb roots, which explains the failure rate.

3.3.2.3 Evaluation Results

At this stage of our research, 60% of all the words in our parallel texts can be handled by the united lexicon of *Mvyut* and *TBTY*. Since 40% fail on account of insufficiency of word items, our main purpose will be to increase the lexical items available to us.

There are a number of other approaches we need to take. We have suggested derivative relations in Sanskrit may be very useful if available. Since words derived from the same word may have meanings that differ from each other, despite certain similarities it is imprudent to treat all the words derived from the same word as if they are are exactly the same. We hold that it is attractive to take into account the way the word derivatives influence the word sense. We hope we will be able to trace fine differences of the sort of word senses as seen in the derivatives by means of existing parallel lexicons.

4. Grammatical Information

When we treat Sanskrit texts, the word inflection and sandhi are quite troublesome. Therefore we are collecting grammatical information, concerning the word inflection. In this section, the collection and evaluation of this kind of information will be the main topic.

4.1 On Accumulating the Information

4.1.1 Nouns

Prior to accumulating noun information, we want to know the distribution of words – which gives us the probability of the occurrence of each stem type. Since appropriate references are not available, we pay attention to the index of *SDP*[5)], which has detailed source information for each word. Relying on this, that is to say, by taking up the source information regarding each word in the index of *SDP*. We can acquire the approximate probability of occurrence of each stem type. However, we picked up some of the pages at random and listed occurrences, since it is impossible to count up all the source information in the index to *SDP*, our interim result from 100 pages of the index to *SDP* is shown in Table 5.

Table 5: Distribution of Words in Sanskrit

Occurrence		Category
2819	(47.8%)	N/a
2059	(34.9%)	P
268	(4.5%)	N/i
197	(3.3%)	V
125	(2.1%)	N/man
112	(1.9%)	N/A
96	(1.6%)	N/u
78	(1.3%)	N/I
50	(1.8%)	N/s
36	(0.6%)	N/in
		All: 5900

In Table 5, *N* indicates the noun and the following symbols signifies the types of word stem, for example, "*N/a*" means the "noun a-stems". Similarly, "*V*" indicates verbs and "*P*" the pronouns. From this table, we can say that about half of all the words in Sanskrit are a-stem noun word. On the other hand, it is demonstrated that the probabilities of verb occurrences are comparatively low.

4.1.2 Verbs

Verbs are very important, because they are the most primitive of Sanskrit word forms. As mentioned above nouns derive from verb roots. But the rules of derivation from verb roots are too diverse and complex for us to process on a computer at present.

We are collecting correspondence tables for verbal roots and stems, along with lists of terminations. We have collected present, aorist, perfect and future stems[5]. Our process is :

- pick up verb roots from Takashima[12)]
- adapt the word form, especially separating preverbs from verb root. For example, change "anubhU" into "anu%bhU". We modified 1997 word forms of verbs.
- generate various verb stems from roots automatically as well as manually. following the rules made by Tsuji[13)] for automatic generation.
- digitalize the list of terminations.
- adapt the information of internal sandhi, which activate when verb stems and terminations are united.

Table 6: Distributions of Verbs in Sanskrit

Occurrence		Conjugation
85	(43.1%)	present
36	(18.3%)	absolutive
24	(12.2%)	optative
18	(9.1%)	present/causative
7	(3.6%)	absolutive/causative
7	(3.6%)	optative/causative
7	(3.6%)	imperative
5	(2.5%)	infinitive
4	(2.0%)	future
		All: 197

To specify Table 5, we investigate the detailed distributions of verbal conjugation, as shown in Table 6. We are still in the process of improving the information for the present system which needs the optative, imperative and imperfect, the future system awaits. It will be ideal if we have information to enable us to parse 60% verbs found in Sanskrit texts automatically. If the results are less than 60%, we must examine what is insufficient for parsing.

4.2 Evaluation

Table 7: Result of Word Parsing

	Stem	Parsed		All
Noun		629	(39%)	1605
	a	497	(40%)	1238
	A	34	(41%)	82
	i	34	(60%)	56
	vat	26	(59%)	44
	u	38	(80%)	47
Verb		95	(39%)	241

We implemented a simple word-parser for the evaluation of the information we heve collected. Our parser largely relies on Takashima[12)] in terms of verbs as well as nouns. On the input of a noun, our system first searches for possible terminations, and then for the entries in Takashima's dictionary (in case of verbs, for the verb stems in the table) and finally decides whether proper noun stems exist. If the system can not find proper noun stems, parsing ends up unsuccessfully.

We select the 22nd chapter of *SDP* for the experiment as in section 3. Since each word in this text has information on its stem, we can easily evaluate parsing results which are shown in Table 7.

Even in the noun a-stem system, parsing accuracy is not high. The reason is that the 532 words of the failured 741 a-stems in *SDP* were not found in Takashima (which includes 40424 items). Among them are some of words frequently found in Buddhist texts, such as "bodhisattva" and "tathAgata". We assume the greater part of the other 209 a-stems are influenced by external sandhi. In Sanskrit, coalescences of words sometimes cause changes in the final letter and the initial letter of the following word. This is known as "external sandhi". For example, "antikAt", which

[5] This process is detailed in 3), 4).

can be parsed by our system, changes to "antikAc" when the following word of a sentence begins with the letter "c". Our system can not parse the new form "antikAc" at present. Our next step is to collect such information on the external sandhi and find out an efficient way to apply it.

For verbs, parsing accuracy resembles that of nouns. Details of 146 failures are shown in Table 8. In this Table 8, Dictionary indicates the number of failures caused by the insufficiency of Takashima. In this Dictionary, problems are divided into Item and Grammar. We refer to information concerning the conjugation for each verb stem which is described in Takashima. 5 failures in Grammar were caused by the insufficiency of this grammatical information in Takashima. Sandhi denotes the failures caused by external sandhi.

Table 8: Result of Verb Parsing

Reason		Failure
Dictionary		40 (27%)
	Item	35 (24%)
	Grammar	5 (3%)
Verb Derivative		83 (57%)
	Absolutive	49 (34%)
	Causative	11 (8%)
	Periph. Pf.	8 (5%)
	Infinitive	5 (3%)
Others		23 (16%)
	Sandhi	14 (10%)
		All: 146

Verb Derivative in Table 8 is also due to the limitation of our information. For example, absolutives failed in 49 words, infinitives failed in 5 words, etc. This insufficiency is caused by the fact that only four verbal stems are supplied. As it turns out, this level of information is insufficient and must be improved and derivatives added.

5 Conclusion

Our purpose is to construct a computer environment in which both languages can be handled seamlessly in terms of e-texts. This project is still in its early stages, and we are now collecting relevant information that has already been digitalized in order to confirm what linguistic information will be desirable.

In this article, we have focussed on two types of data. The first is parallel texts of Sanskrit and Tibetan, useful for researchers. The second is information extracted from original texts, which are needed for computer processing, including the correspondence tables for verbal roots and stems in Sanskrit, and the parallel lexica for Sanskrit and Tibetan.

We have confirmed the usability of the parallel lexicons. If we merge them, 60% of a Buddhist text in both Sanskrit and Tibetan can be covered even at the present stage of the project. To improve this coverage drastically, we must prepare a large quantity of information. Another way to improve the lexicon is to focus on the derivatives in Sanskrit.

In the fourth section, we have investigated the usability of grammatical information. Though parsing accuracy is not good, we have discovered what kind of information is needed. We exhibit our parsing system via WWW[2)].

We are now improving our search system for the parallel texts. We hope our new system employing lexica and grammatical information will stimulate new ideas.

One of the important tasks for the future is to try to improve the parallel texts in Sanskrit and Tibetan. Since statistical research needs plenty of information, improving the quantity of the texts increases its usability enormously. Research concerning the parallel texts continues. Following this trend, studies of bilingual text alignment are also very active. Unfortunately, we can't immediately bring these results into our research because the information needs to brought in to a computer usable format.

Our current concern has focused solely on the Indian Buddhist texts. However, Buddhism has a long history, influencing the culture and society of very wide areas of Asia. Our future hope is to process and relate a variety of texts in various classical languages, including Chinese, Mongolian, and Japanese.

References

1) T. Aiba. Table of Transliteration Schemes for Sanskrit and Tibetan. WWW. (May 20, 2002) URL: <http://texa.human.is.tohoku.ac.jp/aiba/codes/ table/>.

2) T. Aiba. Verb Analyzer for Sanskrit. WWW, 2002. (Aug. 6, 2002) URL: <http://www-asia.human.is.tohoku.ac.jp/demo/ vasia/html/>.

3) T. Aiba and K. Oide. Automatic morpheme analysis of verbs in classical sanskrit: An attempt with the present conjugation. *IPSJ SIG Notes*, 2001-CH-49-1:1–8, 2001. (in Japanese).

4) T. Aiba and K. Oide. Automatic morpheme analysis of verbs in classical sanskrit: Towards an attempt constructing a morphemic dictionary of verbs. *IPSJ SIG Notes*, 2002-CH-53-5:33–40, 2002. (in Japanese).

5) Y. Ejima, editor. *Index to the Sad-dharmapuNDarIkasUtra — Sanskrit, Tibetan, Chinese —*. the Reiyukai, Tokyo, 1985.

6) H. Mitsuhara. Digitalized text of *MahAvyutpatti* based on Sakaki's edition. WWW. (May. 1, 2002) URL: <http://texa.human.is.tohoku.ac.jp/aiba/archive/ mvyut/open/>.

7) K. Nagasaki. Mojiyaki: Multilingual Imager for Web. WWW. (May 20, 2002) URL: <http://mojiyaki.aa.tufs.ac.jp/>.

8) R. Sakaki. *MahAvyutpatti*. Kyoto, 1916.

9) T. Suzuki. project eDic. WWW. (Dec. 20, 2001) URL: http://www.fis.ypu.jp/~suzuki/edic/ (in Japanese).

10) T. Suzuki. "SuvarNaprabhAsa TathAgatAyuHpramANanirdezaparivarta" and "MahAmeghasUtra". *the Bulletin of Institute of Oriental Culture, University of Tokyo*, 135:1–48, 1998. (in Japanese).

11) T. Suzuki, T. Aiba, and M. Matsumoto. Towards the Construction of eDic. URL: <http://texa.human.is.tohoku.ac.jp/aiba/project/ edic/pr/document/genko2.ps> (in Japanese), Jan. 2000.

12) J. Takashima. Sanskrit lexical database based on the Practical Sanskrit Dictionary of V. S. Apte, version 1.0beta. WWW, 2000. (Dec. 1, 2000) URL: <http://www3.aa.tufs.ac.jp/~tjun/sktdic/>.

13) N. Tsuji. *Sanskrit Grammar*. Iwanami Shoten, Tokyo, 1974. (in Japanese).

14) U. Wogihara and C. Tsuchida. *Sad-dharmapuNDarIka-sUtram*. The Seigo-Kenkyukai, Tokyo, 1934.

15) K. Yokoyama and T. Hirosawa. Chinese-sanskrit-tibetan table of buddhist terminology based on the *yogacArabhUmi*. WWW. (Aug. 1, 2000) URL: <http://www.buddhist-term.org/yoga-table/>.

16) K. Yokoyama and T. Hirosawa. *Index to the YogacArabhUmi (Chinese-Sanskrit-Tibetan)*. Sankibo Busshorin Publishing, Tokyo, 1996.