人文科学とコンピュータ　53－5
（２００２．１．２６）

古典サンスクリット語の動詞解析用データの構築とその応用

相場徹　　生出恭治
{aiba,k-oide}@human.is.tohoku.ac.jp
東北大学 大学院情報科学研究科

概要: 古典サンスクリット語 (Skt.) 文法は複雑なため、Skt. の電子テキストに対して計算機を用いて高度な解析を行うことは現状では困難である。我々は、高度な解析を行うための最初のステップとして、Skt. 動詞解析の可能性について検討している。

我々はまず動詞語幹および人称語尾をカテゴリ化し、形態素辞書を構築していくことから始めた。既存の形態素解析システムの流用を考慮に入れ、辞書のデータ形式は日本語の書式に従った。連声の規則は、表を作成することによって対処することとした。

現状における我々の成果はまだ Skt. を自由に扱うまでには達していない。それゆえ、今後もデータを改善し続ける必要がある。

Automatic Morpheme Analysis of Verbs in Classical Sanskrit: Towards an Attempt Constructing a Morphemic Dictionary of Verbs

Tooru AIBA　　Kyoji OIDE
{aiba,k-oide}@human.is.tohoku.ac.jp
Graduate School of Information Sciences,
Tohoku University

Abstract: The extensive use of prefixes and suffixes in the grammar of classical Sanskrit makes it difficult to undertake an advanced analysis of Sanskrit E-texts by computer. This paper attempts an automatic analysis of verb morphemes in Sanskrit as a possible first step towards a more advanced analysis.

We begin by categorizing verb stems and endings in order to construct a dictionary of verb morphemes. This dictionary is created in order to comply with the requirements of an existing Japanese word parsing software program on which we initially intend to rely. Analytical difficulties caused by the phonetic changes following the principle of sandhi, are handled by the use of a table.

Current results are far from sufficient, but are enough to suggest that improvement in the underlying grammatical data may provide the clues to obtain more complete explanations.

1 はじめに

筆者らは Tibetan-Sanskrit 構文対照電子辞書プロジェクト eDic [7, 8] に参加している。インド仏教の研究においては、一次資料であるサンスクリット語原典の多くが散逸しているため、チベット語訳・漢訳などの翻訳資料を用いた異訳対照比較を通じて、散逸した原典のありうべき内容の推定をおこないながら研究を進める必要がある。eDic は、現存するサンスクリット語原典とそのチベット語訳の文レベルの対応箇所を示すという、インド仏教学研究の方法論に密着したツールの提供を目的としたプロジェクトである。

このプロジェクトは、現在のところ、仏教系テキストである *Saddharmapuṇḍarīka (SDP)* のサンスクリット語原文 [13] と、そのチベット語訳 (ラサ版) について、手作業によって、電子化および図 1 のような書式での対訳データの整理をおこなっている。また、すでに入力済の一部テキストについては、チベット語の単語をキーワードにしたテキスト検索サービスをおこなっている。しかし、サンスクリット語をキーにした検索については、現在のところ、単語の語形表現の多様性に対応する方法が確立されていないため、検索サービスを行うことができない。また eDic 本来の目的を考えると、将来的にはサンスクリット語・チベット語のそれぞれを用いた類似文・平行文の提示など、何らかの知識処理を行っていく必要がある。

このような状況において、筆者らは古典アジア系諸言語、なかでも古典サンスクリット語を計算機処理するための枠組の必要性を感じるに至った。そしてその最初の段階として、計算機による単語解析処理の問題に着手した。

Sanskrit (22.2.15–22.2.16)
sa ca sarvasattvapriyadarśano **bodhisattvo** mahāsattvas tasya bhagavataḥ pravacane duṣkara-caryā abhiyukto abhūt
Tibetan (237b3–237b4)
byaṅ chub sems dpaḥ = sems dpaḥ chen po = sems can thams cad kyis mthoṅ na dgaḥ ba = de - ḥaṅ/ bcom ldan ḥdas = de - ḥi = gsuṅ rab - la = dkaḥ ba byed pa - la = brtson pa - r = gyur te/

図 1: eDic のデータ例

原 文

sukhaviśiṣṭasya prakṛtasyaiva brahmaṇotrābhidhānādayamakṣyādhāraḥ paramātmā/

⇓

単語に分割

sukha/viśiṣṭasya prakṛtasya/eva brahmaṇas/atra/-abhidhānāt/ayam/akṣi/ādhāras parama/ātmā/

⇓

接辞処理

sukha/viśiṣṭ-asya prakṛt-asya/eva brah-maṇas/atra/abhidhān-āt/ayam/akṣi/ādhār-as parama/ātmā/

⇓

単語解析

sukha=viśiṣṭa(sg. G.) prakṛta(sg. G.) eva brahman-(sg. G.) atra(L.) abhidhāna(sg. Ab.) / idam(sg.Ab.) akṣi=ādhāra(sg. N.) parama=ātman(sg. N.)

図 2: サンスクリット文の単語処理

2 古典サンスクリット語

2.1 語解析の本来的な枠組

サンスクリット文の一例 [2, p.253] を図 2 に示す。ここで、原文は空白記号によって 4 つの部分に分割されている。このような原文に対して、単語分割、接辞処理、単語解析を、順にあるいは同時におこなうことによって文の読解が可能となる。この原文の最初にある “sukhaviśiṣṭasya” は 2 つの名詞語幹による名詞複合語 (Nominal compounds) の “sukhaviśiṣṭa” が格変化を起こしたものである。また原文の 3 つめの部分は、最後の “akṣyādhāra” が名詞複合語となっていて、それが他の “brahmaṇas” “atra” 等の単語と、空白をおかずに接続している。

また、単語語形解析とは直接の関連はないが、この文は空白によって区切られた 3 つ目の要素の途中にある “abhidhāranāt” までの部分が副文になっており、単語区切りの有無と、文構造との間には特別な相関が見いだせないことも古典サンスクリット文の特徴としてあげられよう。

2.2 計算機解析

古典サンスクリット語文解析システムについては、Verboom[12] によって試みられたことがあった。また近年においては、インドの Centre for Development of Advanced Computing (C-DAC) による DESIKA というシステムが “Technology Development for Indian Languages (TDIL)”[6] から無償で

入手できる。ただし DESIKA の現在公開されている試作版は、単語の名詞語幹を入力すると、それに対応した語形変化を出力する程度の機能しか持っていない。また、このシステムは独自の GUI インターフェイスを用いた Microsoft Windows95 用のシステムとして公開されているため、何らかのシステムのサブルーチンとしての使い方は、現状では困難である。それゆえ、現状では DESIKA を用いることは不可能と判断した。そして、同じ TDIL のサイトによれば、Academy of Sanskrit Research, Melkote, India においてサンスクリット文の構文・意味解析をおこなうシステムが開発されているとのことであるが、その詳細等については現在調査中である。また最近になって、Huét がサンスクリット語の言語解析用のシステムを用意しつつあり[1]、その成果の一部が 2002 年春頃に公開される見通しであるとのことである [5]。

このような状況の中、高島らによって構築された古典サンスクリット語電子辞書 [9, 10] は見出し語が約 5 万と、豊富な語彙量を持っている。また、古典サンスクリット語以外の言語、たとえば日本語においては単語の形態素解析システムはかなり高い水準に達していることが言われており、たとえば「茶筅」[1] は有用なツールとして広く利用されている。筆者らは、高島らの辞書を用いて独自の形式での動詞語幹辞書の構築を行っているが、この辞書形式を「茶筅」等のシステムで用いられるデータ形式と類似したものとすることによって、将来、本格的な古典サンスクリット語形態素解析システムが構築される契機となるのではないかと考えた。

2.3 文解析のイメージ

古典サンスクリット語はインド・ヨーロッパ語族に属する言語ではあるが、単語間に切れ目を入れない表記が許容されているため、英語等で用いられる接辞処理の手法は基本的には使うことができず、日本語等における形態素解析と同様に、単語区切りを入れる処理を行っていく必要がある。また、単語語形解析の際には、古典サンスクリット語に特有の問題もある。

[1]2001 年 10 月に開催された "Workshop on Computational Linguistics in South Asian Languages" において行った 'Computational Tools for Sanskrit' という講演で、その旨の発表を行ったようである [4]。

実際に、古典サンスクリット語における語形解析システムの構築の際に問題となるのは以下であろう。

1. 単語区切りの際に生じる問題
   - 長大な単語列
   - 単語間の連声 (sandhi)
2. 単語語形解析の際に生じる問題
   - 動詞語解析は語根 (root) までの解析が必要
   - 単語内の連声 (sandhi)

2.3.1 単語区切りの際の問題

長大な単語列 散文においては、相当な数の語幹要素が含まれる単語列あるいは複合語列が作成されやすい傾向にある。それゆえ、単語区切り解析における曖昧性爆発への対処法を講じる必要がある。

単語間の連声 図 2 の原文にある "prakṛtasyaiva" は "-sya" という単語の末尾と、"eva" という単語とが接続する際に起きた音韻変化の結果である[2]。このような音変化は、複合語においても発生し、同じく図 2 にある "akṣyādhāraḥ" は "akṣi" と "ādhāra" という名詞語幹が接続した結果である。このような、単語の接続によって発生する文字変化を「外連声 (external sandhi)」と呼ぶ。このような音韻変化が、古典サンスクリット文の解析を困難にしている原因の一つである。

2.3.2 単語語形解析の際に生じる問題

動詞語解析は語根 (root) までの解析が必要 古典サンスクリット語においては、名詞・形容詞等は語幹がそのまま辞書見出しとされるのに対し、動詞については、語幹 (stem) ではなく、語根 (root) が辞書見出しとして用いられる。語根と語幹の関係について、図 3 に例を示す。このうち "P. 3. sg." で示された "dogdhi" 等が、実際に文中に出現する単語となるが、これらの単語を語幹・語尾とに解析したのち、さらに語幹から語根情報を出すことが動詞語解析には求められる。

[2]「Sandhi の現象はいずれの国語にも多かれ少かれ認められるが、Skt. においては音変化の結果をそのまま文字に書き表すことを常とするところに特徴がある」[11, p.15]

duh (2.P)		
System	**stem**-endings	P. 3. sg.
現在 (2)	**doh**-ti	dogdhi
アオリスト (III)	a-**dūduh**-at	adūduhat
アオリスト (VII)	a-**duh**-sat	adhukṣat
完了	**dudoh**-a	dudoha
未来	**doh**-syati	dhokṣyati
śru (5.P)		
System	**stem**-endings	P. 3. sg.
現在 (5)	**śṛṇo**-ti	śṛṇoti
アオリスト (IV)	a-**śrau**-sīt	aśrauṣīt
完了	**śuśrāu**-a	śuśrāva
未来	**śro**-syati	śroṣyati

図 3: サンスクリット語動詞語根と各種語幹

ただし、語根からの語幹の生成規則については、さまざまな規則および多くの例外が存在しているため、解析された語幹から、何らかの規則に基づいて語根を復元することは不可能と思われる。

単語内の連声 (sandhi) 先に、単語と単語の接続の際に生じる外連声について述べたが、単語内要素、すなわち語幹と語尾等が接続する際にも、内連声 (internal sandhi) と呼ばれる音韻変化が起こる。図 3 の "doh-ti" が "dogdhi" となる等はその典型例である。内連声は外連声と音韻規則が若干異なっている点にも注意する必要があろう。

これらの諸課題のうち、語根と各種語幹の対応の問題については、語根からの各種語幹生成規則に基づいて動詞語根と各種語幹の対応表をあらかじめ作成してしまい、動詞語幹の形態素データに語根情報を埋め込むのが最も現実的な対処法であろう。連声への対処については、連声規則に関する一覧表をあらかじめ作成しておき、語形解析の際には、解析を行う前にプリプロセッサのような形で連声の対処を行う方法が妥当と考えた。作業はこの方針に従って行う。

2.4 データ構築の枠組

前述したとおり、筆者らが当面扱うのは eDic プロジェクトで用意されたデータである。このデータは、一般的な古典サンスクリット語文とは異なり、すべての単語の間にはあらかじめ区切りが入れられている。図 1 の例では、サンスクリット部分の最下行にある "duṣkara-caryā" におけるハイフン部分が、手作業によって入れられた区切り箇所である。

文字組		連声後	文字組		連声後
gh + t	→	gdh	gh + th	→	gdh
ḍh + th	→	ḍḍh	dh + t	→	ddh
ch + s	→	kṣ	kṣ + s	→	kṣ

図 4: 内連声の表

このようなデータを扱うため、前節で述べた古典サンスクリット文の語解析における諸課題のうち、「単語区切りの際の問題」については、当面の課題として扱う必要がない。ただし、長期的な視点からは解決していかねばならない問題であることは間違いないことから、将来の単語区切り処理も視野に入れつつ、当面の課題を克服するための作業を行っていく。

3 内連声への対応

Verboom は BOBRA というプログラミング言語が持つ、強力なパターンマッチ機能を利用して連声を展開させたようである。また Huét が開発しつつあるシステムは有限オートマトンを用いて連声の解析を行うようであるが、具体的な解析の方法についてはまだ不明である。

筆者らが想定する連声への対処の枠組みは、Verboom と同様に、単語解析に先立つプリプロセッサとして機能させるものである。このような対処をおこなうために、連声前・連声後の対応表の作成を行う。

3.1 内連声表の作成

辻 [11] における内連声の規則等を参考にし、図 4 のような一覧を作成する。これは "gh" と "t" とが接続すると "gdh" に変換される、といった類の情報である。このような一覧を作成したのち、今度は、文中語がどのような文字の組み合わせによってできた可能性があるかを知るための一覧を作成する。こうして作成した一覧を図 5 に示す。この一覧を用いると、文中にある "gdh" という文字列は "gh+t", "gh+th" ...のいずれかの文字組が変化した結果である可能性がある、ということがわかる。現在のと

文中		構成要素候補
kṣ	←	k+s, c+s, śc+s, ch+s, j+s, ...
gdh	←	gh+t, gh+th, h+t, h+th, h+dh
ḍḍh	←	ḍh+t, ḍh+th, ṭh+d, ṭh+dh, ...
jñ	←	j+n
antu	←	am+tu
ṭḍh	←	ṭ+dh
nv	←	m+v
ṭṭh	←	ṭ+th, ṭh+t, ṭh+th, ḍ+th, ...
ṝḍh	←	ih+t, ih+th, ih+dh, ṝh+t, ...

図 5: 内連声の逆表

ころ、図 5 に示したような語情報を 500 個程度用意した。

3.2 文字の欠落等への対処

内連声の中では、連声として扱うと処理が複雑になってしまう可能性を持つものがある。

- 文字の欠落が起こる場合の一部
  ("ataut-stam" が "atauttam" になる等)
- 文字の連接箇所とは離れた箇所に影響が及んでいる場合の一部

内連声に対する処理は図 5 に示す一覧を用いて行うが、文字の連接箇所とは離れた場所での音韻変化にまで対応しようとすると、表を複雑にしなければならなくなり、なるべく避けたい。

この後者の代表的なものとして以下が上げられる。

- 語根が g,d,b (無気音) で始まり gh,dh,bh,h (有気音) で終わるときに、語根末尾の有気音が無気音に変わったときは語根先頭が有気音になる
- "n" と "s" の反舌音化 (Cerebralisation)

この前者の代表例が "**d**āh-sam" が "**dh**ākṣam" になるものである。この例は一部特定の単語のみで起こるものなので、現状では「動詞 dah の P. 3. sg. のアオリスト VII 語幹は dhak-」のように、語幹生成時にあらかじめ有気音に関する処理をしてしまうようにしている。

一方、図 6 に示した条件において発生する「"n" と "s" の反舌音化」については、かなり頻繁に見ることができる。この規則は、たとえば "kar-ana" という接続が生じたときに、この "n" は "r" に先行され、先行する "r" との間には母音 "a" のみが介在し、後続する音 "a" は母音であることから図 6 の "n → ṇ" の条件を満たすことになり、その結果 "karaṇa" と音が変化する規則である。

先行音	介在可能な音		後続する音
ṛ, ṝ, r, ṣ	母音, k, kh, g, gh, ṅ, p, ph, b, bh, m, y, v, h, ṃ	**n** ⇓ **ṇ**	母音, n, m, y, v
母音 (a,ā 以外), k, r, (l)	ṃ, ḥ	**s** ⇓ **ṣ**	母音 (ṛ 以外), t, th, n, m, y, v

図 6: n/s の反舌音化 (Cerebralisation)

この規則を完全に満たすためには、相当数の文字に対するパターンマッチを行う必要があるが、筆者らはとりあえずこの規則を、"n" および "s" に先行する文字は見ずに、「母音,n,m,y,v の前にある ṇ は n の可能性がある」「(ṛ 以外の) 母音,t,th ... の前にある ṣ は s の可能性がある」と読み替えて、当面は形態素辞書に両者を併記して対処することとした。

4 動詞語幹・語尾データ構築

古典サンスクリット語の単語の曲用・活用の体系は、大きく名詞・形容詞型のものと動詞型のものに分類される。筆者らは現在、動詞型活用への対処のため、辞書から取り出すことのできる動詞語根から、計算機解析用データとしての各種動詞語幹の一覧を作成している。

作業としては、現在のところ、動詞現在語幹組織およびアオリスト語幹組織の整備がひととおり終了したところである。以下に、動詞語幹や人称語尾をどのように形態素データ化したかについて、順に述べる。

4.1 現在語幹組織

動詞語根からの現在語幹組織辞書の構築については、別稿 [3] において述べている。本節では、このように構築されたデータに、どのような形態素辞書情報を与えるかについて述べる。

動詞現在組織は、動詞語根からの現在語幹生成規則等により、第 1 種活用 (第 1,4,6,10 類および de-

nominative) と第 2 種活用 (第 2,3,5,7,8,9 類) とに大きく分類される。

4.1.1 第 1 種活用

高島 [9] によると、第 1 種活用に属する動詞語根の数は 3,738 個で、動詞見出し全体の約 80 % を占めている。ここに属する第 1,4,6,10 類の動詞語幹はそれぞれ語根から語幹への生成規則が異なっているが、語幹と人称語尾の処理は一括で行うことが可能である。それゆえ、図 7 で示すような記述を行った。

動詞	
語幹:	**bhar**
語根:	**bhṛ**
活用:	現在語幹・第 1 種 (1)・A
動詞語尾	
活用:	現在組織・第 1 種・A
語幹:	*
語尾:	pres. 1. sg. → **e**
	pres. 1. pl. → **āmahe**

図 7: 現在動詞 (1) の形態素辞書記述例

4.1.2 第 2 種活用

第 2 種活用は、活用の形態が類ごとに異なる。また、活用形によって強語幹と弱語幹とを使い分けることがあるため、場合によっては語幹ごとに区別する必要がある。

第 5 類 語根に “no” を添えて強語幹が、“nu” を添えて弱語幹が作られる。それゆえ、添字の “n” までを語幹末尾にすると、強語幹と弱語幹の区別が不要となる。ただし、語根末尾が母音の場合と子音の場合とで活用形が若干異なるため、その両者を区別する必要がある。母音型に関する形態素データの記述例を図 8 に示す。

第 8,9 類 第 8 類、第 9 類ともに、第 5 類とほぼ同じ方法での辞書記述が可能。ただし語根 “kṛ” は pres. 3. sg. P. で “kar-oti”(強語幹)、pres. 3. pl. P. で “kur-vanti”(弱語幹) となり、活用形によって語幹部分が異なってしまう。それゆえ、語幹ごとに別項目として辞書見出しを立てることとした。この辞書記述例を図 9 に示す。

動詞	
語幹:	**sun**
語根:	**su**
活用:	現在語幹・第 2 種 (5) 母音型・P
動詞語尾	
活用:	現在組織・第 2 種 (5) 母音型・P
語幹:	*
語尾:	pres. 1. sg. → **omi**
	pres. 1. pl. → **umas/mas**

図 8: 現在動詞 (5) の形態素辞書記述例

動詞	
語幹:	**kar**
語根:	**kṛ**
活用:	現在語幹・第 2 種 (8) kṛ型 強・P
語幹:	**kur**
語根:	**kṛ**
活用:	現在語幹・第 2 種 (8) kṛ型 弱・P
動詞語尾	
活用:	現在組織・第 2 種 (8) kṛ型 強・P
語幹:	kar
語尾:	pres. 1. sg. → **omi**
	pres. 1. pl. → ×
活用:	現在組織・第 2 種 (8) kṛ型 弱・P
語幹:	kur
語尾:	pres. 1. sg. → ×
	pres. 1. pl. → **mas**

図 9: 現在動詞 (8) (kṛ) の形態素辞書記述例

第 2,3,7 類 いずれも動詞語根に直接人称語尾が接続するため、内連声に関連した例外的活用が多い。なかでも第 2 類の動詞語根には特殊な活用をするものが多く、“īś”, “brū”, “yu”, “stu”, “duh” など、かなりの数の動詞を区別する必要があった。ただし、これらのうちの一部は、内連声の処理をきちんと行うことができるようになれば例外として扱う必要がなくなると考えている。これについては今後も課題となろう。

4.2 アオリスト語幹組織

動詞アオリスト組織は、動詞語根からの語幹の生成規則に基づき、図 10 に示すとおり 7 種類に分けられる。これらは、さらに単純アオリストと歯擦音アオリストとに大別される。

type	root	stem	3. sg. P.
I	dā	dā-	adāt
II	sic	sic-	asicat
III	śri	śiśriy-	aśiśriyat
IV	nī	nai-	anaiṣīt
V	lū	lāv-	alāvīt
VI	yā	yā-	ayāsīt
VII	diś	diś-	adikṣat

図 10: アオリスト語幹

動詞	
語幹:	**d**
語根:	**dā**
活用:	Aor. 語幹・I・P
語幹:	**bh**
語根:	**bhū**
活用:	Aor. 語幹・I bhū 型・P
動詞語尾	
活用:	Aor. 組織・I・P
語幹:	*
語尾:	1. sg. → **ām**
	1. pl. → **āma**
活用:	Aor. 組織・I bhū 型・P
語幹:	bh
語尾:	1. sg. → **ūvam**
	1. pl. → **ūt**

図 11: アオリスト動詞 (I) の形態素辞書記述例

4.2.1 単純アオリスト

I(Root-aor.), II(a-aor.), III(Reduplicated aor.) は、それぞれ語幹生成規則と人称語尾体系が異なっているが、いずれも語幹と語尾が接続する場所に母音が入っているという点で共通している。それゆえ母音以降を人称語尾、その直前の子音までが語幹、という形で語幹部分と語尾部分とを切り分け、図 11 に示すとおりに形態素データ化する。これらの中では、語根 "bhū" のみが特殊な人称語尾を取るため、他の活用とは分けて記述している。

4.2.2 歯擦音アオリスト

語根部と語尾の間に歯擦音 "s" を挟むのが歯擦音アオリストの特徴となっている。本節では、これらのアオリストの中から、複雑な対処が必要となる IV, V の形式についてのみ述べる。

V: iṣ-aor. (8) の動詞は、2.3. sg. A. で任意に鼻音を省くことができる点が問題となる。語根 "tan"

	nī	**tud**	**dah**
sg.1.	anaiṣam	atautsam	adhākṣam
2.	anaiṣīs	atautsīs	adhākṣīs
3.	anaiṣīt	atautsīt	adhākṣīt
du.1.	anaiṣva	atautsva	adhākṣva
2.	anaiṣṭam	atauttam	adāgdham
3.	anaiṣṭām	atauttām	adāgdhām
pl.1.	anaiṣma	atautsma	adhākṣma
2.	anaiṣṭa	atautta	adāgdha
3.	anaiṣus	atautsus	adhākṣus

図 12: Aorist IV (s-aor) 活用例

は 2. sg. A. で "**atan**iṣṭhās" となるのが普通であるが、"**ata**thās" という形式も任意に取り得るようである。この場合は別語幹・別活用語尾として区別して扱う。

IV: s-aor. この活用例を図 12 に示したが、"nī" のように、人称語尾の先頭に "s" が挿入されるのが特徴である。図 12 の "tud", "dah" のように、語幹末尾が子音の時などは一部活用において挿入文字 "s" が脱落する。また "dah" のように語根が "h" で終わるものは、一部活用において語幹部分が "**dā**g" ではなく "**dhā**k" となる。これについては、当面は別語幹という扱いでの対処を行う。

5 おわりに

筆者らは、古典サンスクリット動詞解析を目的として、動詞解析用データの構築を行った。しかし、古典サンスクリット語の文法規則は非常に複雑で、例外も非常に多い。それゆえ、形態素データおよび連声の規則の有効性を確認するため、筆者らは簡単な語形解析プログラムを作成し、さまざまな語形がきちんと解析されるか否かに関する確認を行いながら、データの有効性向上に取り組んでいる。

サンスクリット語の辞書においては、動詞語根見出しとともに、その動詞語根が取り得る各種語幹が掲載されていることがある。しかし、それら各種語幹は 3. sg. の活用形のみを示すことがほとんどであり、それ以外の活用形についての記述がない。それゆえ、辞書等にある記述は、筆者らが構築したデータ全体を評価する目的には使うことができない。このように、解析用データの有用性を図るための環境がまだ十分に整備されていないため、本稿では筆者

らが構築したデータの有用性等に関する評価を行うことはできなかった。筆者らの構築した成果について、きちんとした評価を行うための環境の構築は、今後における重要な課題であると考えている。

数値的な評価ではなく、あくまで筆者自身が受けた印象を述べるならば、本データには以下のような課題があり、今後解決していく必要があると考えている。

動詞語根からのアオリスト語幹の生成　現在語幹については、辞書見出し等に「duh (2)」との記載があるため、きちんとした形態素データの整備が行われていたのに対し、アオリスト語幹は、どの語根がどの規則に従って語幹の生成を行うかが明記されていないため、本来であれば存在しているはずの語幹が現状では相当数落ちているように思われる。この欠落を今後どのように埋めていくかが課題となる。

内連声の対応表の精度向上　筆者らが取った枠組においては、内連声によって音韻変化する可能性のある文字組をあらかじめ展開するものであった。この場合、たとえば「反舌音 (.s を含む) + 歯音 (n を含む) > 反舌音 + 反舌音」というような、ある程度の数の文字が含まれるもの同士の連声規則に対処するためには、それら集団に含まれるすべての要素の組み合わせを用意する必要がある。このような総当たりを取ると、実際には連接のない文字組のデータまで登録されてしまうことになり、無駄な情報量が多くなってしまう。よって、解析に必要な情報を今後も加えていくのは当然であるが、不要な情報の削除も行っていきたい。

形態素辞書におけるデータの書式の検討　現状のデータ形式は、サンスクリット語のデータを記述するための最適な形式であるとは言いがたい。そこで、様々な言語における、形態素辞書の書式に関する調査を行い、書式の改訂を行っていくことも今後の課題となる。

謝辞

本稿で用いた電子辞書について、我々が研究に利用することを許可してくださった東京外国語大学アジア・アフリカ研究所の高島淳先生に感謝いたします。

インド仏教学研究者という立場からの電子データの利用について、さまざまな意見交換をおこなった山口県立大学の鈴木隆泰さん、東北大学大学院文学研究科の松本峰哲さんをはじめとした eDic プロジェクトのメンバーの方々に感謝いたします。

古典サンスクリット語の動詞規則等について貴重なご意見をいただいた東北大学大学院文学研究科の笠松直さんに感謝いたします。

参考文献

[1] 形態素解析システム 茶筌. WWW. (Dec.20, 2001) URL: <http://chasen.aist-nara.ac.jp/>.

[2] V. S. Abhyankar. *Śrî-Bhâshya by Râmânujâcârya.* No. 68 in Bombay Sanskrit and Prakrit Series. Bombay, 1914.

[3] 相場徹, 生出恭治. 古典サンスクリット語動詞現在組織の形態素解析とその問題点. 情報処理学会研究報告, Vol. 2001-CH-49-1, pp. 1–8, 2001.

[4] M. Butt and T. H. King. Workshop on Computational Linguistics in South Asian Languages Tuesday, 9 October, 2001 As part of the XXIth South Asian Languages Analysis Roundtable meeting. WWW. (Dec. 20, 2001) URL: <http://ling.uni-konstanz.de/pages/conferences/sala01/cl-workshop.html>.

[5] G. Huét. Gérard huét's sanskrit site. WWW. (Dec. 20, 2001) URL: <http://cristal.inria.fr/~huet/SKT/>.

[6] Ministry of Information Technology, India. Technology Development for Indian Languages. (Dec. 20, 2001) URL: <http://tdil.mit.gov.in/>.

[7] 鈴木隆泰. Tibetan-Sanskrit 構文対照電子辞書プロジェクト eDic. WWW. (Dec. 20, 2001) URL: <http://www.fis.yamaguchi-pu.ac.jp/~suzuki/edic/>.

[8] 鈴木隆泰, 相場徹, 松本峰哲. Tibetan-Sanskrit 構文対照電子辞書 eDic の構築に向けて. 東京大学東洋文化研究所 1999 年度班研究「インターネット利用技術」研究会用資料, 18 pages, Jan. 2000.

[9] 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/%7Etjun/sktdic/>.

[10] 高島淳. サンスクリット語の機械可読辞書の開発とパーザへの適用. 平成 9 年度〜平成 11 年度科学研究費補助金 基盤研究 (A) (2) 研究成果報告書『インド諸言語のための機械可読辞書とパーザの開発』 (課題番号 09044004) (研究代表者 ペーリ・バースカラーオ), pp. 73–105, 2000.

[11] 辻直四郎. サンスクリット文法. 岩波全書 280. 岩波書店, 東京, 1974.

[12] A. Verboom. Towards a Sanskrit wordparser. *Literary & Linguistic Computing*, Vol. 3, No. 1, pp. 40–44, 1988.

[13] U. Wogihara and C. Tsuchida. *Saddharmapuṇḍarīka-Sūtram.* The Seigo-Kenkyūkai, Tokyo, 1934.