人文科学とコンピュータ 49－1
（2001. 1. 19）

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

相場徹　　生出恭治
{aiba,k-oide}@vacia.is.tohoku.ac.jp
東北大学大学院 情報科学研究科 (片平)
仙台市青葉区片平2丁目1-1

概要: 昨今になって、古典サンスクリット語 (Skt.) の電子テキストが大量に入手できるようになった。しかし電子テキストを解析するための知識の蓄積はまだ十分ではない。そこで本稿では、本格的なテキスト解析へ向けた最初の試みとして、Skt. の動詞現在組織の形態素解析への取り組みについて述べる。

我々はまず動詞語根からの現在語幹の生成に関する文法規則の整理をおこない、その規則に基づいて動詞現在語幹を自動的に生成する。この結果、全体の約9割の動詞語根から正しく動詞現在語幹が生成されることを確認した。もう一つの問題として、人称語尾が語幹に接続する際に「連声」によって生じる音韻変化がある。我々はあらかじめ連声に対応した表を作成することによって、これに対処することとした。

Automatic Morpheme Analysis of Verbs in Classical Sanskrit: An Attempt with the Present Conjugation

Tooru AIBA　　Kyoji OIDE
{aiba,k-oide}@vacia.is.tohoku.ac.jp
Graduate School of Information Sciences at Katahira,
Tohoku University
Katahira, Aoba-ku, Sendai 980-8577, Japan.

Abstract: There are now many more E-texts in classical Sanskrit, but we still cannot easily analyze them using current software tools. In this paper we consider the possibility of automatic morpheme analysis of verbs in present conjugations in classical Sanskrit as a first step towards a more developed analysis.

We begin by categorizing the grammar rules for generating the verbal present stems from the verb-roots. This shows about 10% of verbal present stems are not generated through a consistent application of the rules. Another problem is caused by sandhi, which applies phonetic rules governing both the stem and the endings of verb forms. We deal with this by a table to aid the digital recognition of verb forms.

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 [6, 7] 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。しかし古典サンスクリット語・チベット語に関する計算機的な知識の蓄積がいまだ十分ではなく、文中単語の原形の自動推定もできない現状にあっては、プロジェクトの目的達成への道のりは遠い。

一方、昨今になって古典インド学仏教学関連分野においては、京都大および京都産業大 [11]、JBE [5]、ACIP [2]、SAT [9]、CBETA [3] 等により、かなり大量の電子テキストが構築・公開されるようになってきている。これらで公開されているテキストは、テキストの内容以外の情報 — サンスクリット語であれば文中のそれぞれの語についての品詞・活用・語幹等の情報 — が付与されておらず、単語検索以外の用途で用いることは困難である。それゆえ、電子テキスト中の各単語について、自動的に品詞・活用・語幹等の情報を付けることができれば古典サンスクリット語を計算機で扱う研究が飛躍的に進展する可能性が生じてくる。

このような状況において、筆者らは高島 [8] を閲覧し、研究への利用を許諾される機会を得た。この電子辞書は Apte [1] に基づいたものであり、非常に重要で価値の高いものと考えられる[1]。そこで、この辞書を単語辞書として用い、古典サンスクリット語文献を計算機で扱うための最初のステップとして必要となる、単語の形態素解析に関する研究に取り組むこととした。

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。本稿ではこのうち動詞型のものを取りあげた。辻 [10, p.109] は動詞の活用を以下のような組織 (system) に分類している[2]。

1. 現在組織 (直説法現在、現在分詞、直説法過去、願望法、命令法)
2. アオリスト組織 (直説法アオリスト、祈願法)
3. 完了組織 (直説法完了、完了分詞)
4. 未来組織 (直説法未来、未来分詞、条件法)
5. 複合時制 (複合完了、複合未来)
6. 語根直属の準動詞 (過去分詞、動詞的形容詞、不定詞、絶対分詞)

本稿では、このうち現在組織のみを扱う。現在組織に属する各種活用は活用のパターンが多数あり、別の動詞組織を扱う際にも問題となる「重字」「連声」などへの対処が必要とされているため、動詞現在組織の解析を可能とすることは同時に、他の活用組織の解析にとっての基礎的な技術の蓄積にもなるからである。それゆえ動詞現在組織の解析という課題は、古典サンスクリット語の形態素解析システム構築の実現性をはかる尺度となり得ると考えられる。

ところで辻 [10] においては、随所で「特例」「一般に」「若干の」という単語が用いられており、一般的な文法規則に従わない言語的用例が相当数存在することを予感させる。それゆえ、文法書における文法規定が実際の言語に対してどの程度有効に機能し得るのかに関する見通しを立てることについても、形態素解析システム構築の実現性をはかる点では重要であると思われる。

2 サンスクリット語

2.1 連声と母音の階梯

古典サンスクリット語の動詞現在組織の処理について述べる前に、サンスクリット語処理一般において問題となる連声法 (sandhi) および母音の階梯 (vowel gradation) についてあらかじめ述べておく。

連声 文中または語中における音の連結に関する規則を総称して連声法と呼ぶ。これは、たとえば “tat śrutvā” という単語列が、実際に文中に出現する際には単語の連結部分が音韻的な変化を起こし “tac chrutvā” となるような規則である。古典サンスクリット語の読解の際には、このような単語間の連結

[1]「文学作品を読む時は最も有益なアプテ (V.S.Apte) の梵英辞典」 [4, 上.p.10]

[2]以下、本稿で述べる古典サンスクリット文法に関する記述は辻 [10] を元にしている。

	I	II	III	IV	V
弱 韻	–	i,ī, (y,iy)	u,ū (v,uv)	ṛ (r), ṝ (ir, ur)	ḷ
Guṇa	a	e (ay)	o (av)	ar	al
Vṛddhi	ā	ai (āy)	au (āv)	ār	–

図 1: 母音の階梯

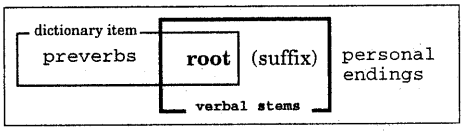

図 2: 動詞の一般的な構造

部分における連声の規則を知らないと単語の認識さえ不可能である。

本稿で述べる枠組においては、この連声法は単語内の各要素の連結の際に問題となる。文中に出現する動詞は、語幹 (stem) 部分に人称語尾 (personal endings) 等が接続して生成されたものであるが、この語幹部分と人称語尾等との接続部分においても連声による音韻変化が起こるのである。たとえば動詞語根 "yuj" から生成される現在語幹 "yunaj" に人称語尾 "si" が接続すると "yunajsi" ではなく "yunakṣi" となる。

母音の階梯 動詞語根 (root) からの動詞語幹の生成においては、母音の階梯に関する知識も必要である。図 1 に母音の階梯の一覧を示す。この図のうち、括弧で示したものは直後に母音が来る場合のものである。たとえば動詞語根 "ji" から動詞現在語幹が生成される場合、語根の母音が guṇa 化したうえで接尾辞 "a" が付けられる規則であるため、図 1 の II の規則により "ji" の "i" が guṇa 化して "ay" となり、それに接尾辞 "a" が接続した結果 "jaya" となる。

2.2 動詞の構造

古典サンスクリット語における現在動詞の構造を図 2 に示す。動詞はまず動詞語根があり、その語根に接尾辞 (suffix) が付くなどにより各種動詞語幹が形成される。そして、その動詞語幹に人称語尾を接続し、また動詞によっては動詞接頭語が付けられる形で動詞が形成される。

upanahyati		
(prev)	(stem)	(ending)
upa -	nahya -	ti
	(nah - ya)	
upa	nah	*(→ dict.)*

図 3: 現在動詞の一例 (1)

pradvekṣi		
(prev)	(stem)	(ending)
pra -	dveṣ -	si
	(dviṣ - ×)	
pra	dviṣ	*(→ dict.)*

図 4: 現在動詞の一例 (2)

図 3 に、現在動詞 "upanahyati" を各要素に分解した例を示す。この単語は、接頭辞・語幹・語尾ごとに "upa-nahya-ti" と区切りを入れることができ、また語幹は語根部分と接尾辞とで "nah-ya" と区切ることができる。そして、接頭辞と語根を組み合わせた "upanah" が辞書の見出し語となる。

"upanahyati" の例では、動詞からの動詞語根の発見は容易であったが、動詞によっては語根部分の推定が困難な場合もある。図 4 に示す "pradvekṣi" という現在動詞の語幹は "dveṣ" で、人称語尾は "si" である。これは動詞語幹と語尾の連結部分に連声法が適用され、"dveṣ-si" が "dvekṣi" となった結果である。また語幹 "dveṣ" の語根は "dviṣ" である。このような場合は動詞の語根推定は容易ではない。

2.3 動詞語根の類別

動詞語根は、現在動詞語幹の生成規則に基づいて「1 類」から「10 類」までの「類 (class)」に分類される。たとえば図 3 の "nah" は 4 類、図 4 の "dviṣ" は 2 類である。これらの類は、さらに「第 1 種活用」「第 2 種活用」という枠組にグループ化される。第 1 種活用には (1),(4),(6),(10) 類が属し、それ以外は第 2 種活用に属する[3]。これら以外にも、名詞起源

[3]本稿では、読者の便宜を考慮し、以後は「第 1 類」の数字には "()" を付与して「第 (1) 類」と表記する。また、「第 1 種活用」の数字については、数字のみで表記する際には "[]" を付与して "[1]" のように表記する。

Cl.	Verbs	Paras.	Ātm.
(1)	2065 (44.5%)	1507 (40.6%)	787 (39.4%)
(10)	614 (13.2%)	567 (15.3%)	431 (21.6%)
(4)	356 (7.7%)	243 (6.5%)	133 (6.7%)
(6)	355 (7.6%)	326 (8.8%)	74 (3.7%)
(Den)	348 (7.5%)	259 (7.0%)	115 (5.8%)
[1]	3738 (80.5%)	2902 (78.2%)	1540 (77.0%)
(2)	249 (5.4%)	195 (5.3%)	71 (3.5%)
(9)	189 (4.1%)	182 (4.9%)	71 (3.5%)
(8)	147 (3.2%)	144 (3.9%)	139 (7.0%)
(5)	127 (2.7%)	123 (3.3%)	54 (2.7%)
(3)	101 (2.2%)	86 (2.3%)	68 (3.4%)
(7)	94 (2.0%)	80 (2.2%)	57 (2.9%)
[2]	907 (19.5%)	810 (21.8%)	460 (23.0%)
All	4645	3712	2000

表 1: 類ごとの動詞項目数 (高島)

class	root		pres. stem
(1)	budh	→	bodha-
(4)	man	→	manya-
(6)	tud	→	tuda-
(10)	cur	→	coraya-

図 5: 第 1 種活用の現在語幹

の動詞 (denominatives, Den.) も辞書には載せられている。Den. の活用規則は第 1 種活用と同じなので、本稿では第 1 種活用に含めて扱う。

表 1 に、高島 [8] における各動詞の類ごとの動詞数を示す。この表によると、第 1 種活用に属する (1),(4),(6),(10),(Den) 類の動詞語根の数が全体の 8 割を占めていることがわかる。

2.3.1 第 1 種活用

図 5 に、第 1 種活用に属する動詞における語根と現在語幹の組の例を示す。これらの類のあいだには現在語幹の生成方法に若干の相違が見られ、たとえば (1),(10) はある一定の条件のもとで語根母音が guṇa 化するのが一般的であるが (4),(6) は guṇa 化しないのが一般的であること、接尾辞が類によって微妙に異なっていること等が相違点としてあげられる。しかし、これらの微妙な差異にさえ注意すれば、現在語幹の生成規則はかなり類似していることもあり、語根からの自動的な現在語幹の生成は行いやすいと考えられる[4]。

また、すべての類で現在語幹の末尾が "a" になる点が共通しており、現在語幹と人称語尾間の連声への対応を統一的に行うことができる。図 6 に、第 1 種活用の人称語尾の一覧を示すが、図 5 のようにして作成された現在語幹に、図 6 で示した人称語尾を添えるだけで[5]動詞が作成可能である。

	P			A		
	sg	du	pl	sg	du	pl
Present						
1.	āmi	āvaḥ	āmaḥ	e	āvahe	āmahe
2.	asi	athaḥ	atha	ase	ethe	adhve
3.	ati	ataḥ	anti	ate	ete	ante
Imperfect						
1.	am	āva	āma	e	āvahi	āmahi
2.	aḥ	atam	ata	athāḥ	ethām	adhvam
3.	at	atām	an	ata	etām	anta
Optative						
1.	eyam	eva	ema	eya	evahi	emahi
2.	eḥ	etam	eta	ethāḥ	eyāthām	edhvam
3.	et	etām	eyuḥ	eta	eyātām	eran
Imperative						
1.	āni	āva	āma	ai	āvahai	āmahai
2.	a	atam	ata	asva	ethām	adhvam
3.	atu	atām	antu	atām	etām	antām

図 6: 第 1 種活動動詞の人称語尾

class	root	→	pres. stem (strong,weak)
(2)	dviṣ	→	dveṣ-, dviṣ-
(3)	bhṛ	→	bibhar-, bibhṛ-
(5)	su	→	suno-, sunu-
(8)	tan	→	tano-, tanu-
(9)	aś	→	aśnā-, aśnī-
(7)	rudh	→	ruṇadh-, rundh-

図 7: 第 2 種活用の現在語幹

2.3.2 第 2 種活用

第 2 種活用に属する各類では、第 1 種活用に見られたような類間の共通性に乏しい。それゆえ、第 1 種活用のように統一的に扱うことは困難である。

[4] ただし、語根からの現在語幹の生成について、一般的な規則に従わない「例外」が相当量存在していることも無視できない。

[5] 人称語尾の先頭の母音と、現在語幹末尾の母音が重なるため、機械処理の際には、単純に現在語幹末尾の母音 "a" を欠落させたのちに、人称語尾との結合をおこなう。

図 7 に、類ごとの語根と現在語幹の対応表を示す。この図では現在語幹に 2 つの型が示されているが、第 2 種活用の現在動詞語幹には強・弱という 2 つの語幹が存在している。以下に、類ごとに注意すべき事項について簡単に述べる[6]。

(5),(8),(9) 類　図 7 に示すとおり現在語幹生成の際に (5) では "no(強)/nu(弱)"、(8) では "o/u"、(9) では "nā/nī" という接尾辞が語幹に接続する。語幹の末尾が母音であるか否かにより、人称語尾との接続に若干の変化が生じる[7]。

(7) 類　語幹末尾の子音の前に挿入辞 (infix) "na/n" が挿入される。図 7 の例では語根 "rudh" の末尾子音 "dh" の前に挿入辞が入り "ru-**na**-dh" となり、さらに連声によって "ru**n**adh" という語幹となる。

(2) 類　語根に人称語尾が直接接続する[8]。

(3) 類　現在語幹生成の際に、語根の一部が重複する (重字)。図 7 の例では語根 "bhṛ" をもとにした "bi" という重複辞が前置された結果、"**bi**bhar" という語幹等が生成されている。語根に接尾辞は付与されない。また (3) 類では、人称語尾の一部が独自の形をとる点に留意する必要がある[9]。

次に、図 8 に第 2 種活用動詞における人称語尾の一覧を示す。太字で示した部分が強語幹を取るもの、それ以外が弱語幹を取るものである。(2),(3),(7) 類は語幹末尾の文字が語根ごとに異なるため、人称語尾との接続の際に連声法による文字変化が起こる点に注意する必要がある。

	P			A		
	sg	du	pl	sg	du	pl
Present						
1.	**mi**	vas	mas	e	vahe	mahe
2.	**si**	thas	tha	se	āthe	dhve
3.	**ti**	tas	anti	te	āte	ate
Imperfect						
1.	**am**	va	ma	i	vahi	mahi
2.	**s**	tam	ta	thās	āthām	dhvam
3.	**t**	tām	an	ta	ātām	ata
Optative						
1.	yām	yāva	yāma	īya	īvahi	īmahi
2.	yās	yātam	yāta	īthās	īyāthām	īdhvam
3.	yāt	yātām	yur	īta	īyātām	īran
Imperative						
1.	**āni**	**āva**	**āma**	**ai**	**āvahai**	**āmahai**
2.	dhi	tam	ta	sva	āthām	dhvam
3.	**tu**	tām	antu	tām	ātām	atām

図 8: 第 2 種活用動詞の人称語尾

3　現在動詞の形態素解析に向けて

3.1　解析の枠組

我々が取る解析の手順を示す。解析に先だって、まず動詞語根と動詞語幹の対応表をあらかじめ作成しておく。そして、解析の際には、入力単語に対する語幹候補部分と人称語尾部分との切り分けをおこない、あらかじめ用意してある表を用いて動詞語根を特定する。このような手順を用いた動詞の形態素解析をおこなうものとする。

それゆえ、現在動詞の形態素解析のためには、以下のような準備が必要となる。

- 単語辞書をどのように利用するかを決めること
- 動詞語根と、そこから生成される現在動詞語幹との対応表を作成すること
- 人称語尾の一覧表を作成すること。また、語幹と人称語尾が接続する際に起こる連声への対処法を用意すること

このそれぞれについて順に述べる。

3.2　単語辞書

図 9 に、本稿で用いている高島 [8] における辞書記述の一部を示す。ここでは動詞のみを例にあげているが、見出し語 (dictionary item) ごとに類・態[10]の情報が書かれ、動詞によってはさらに各種活用形が書かれたうえで、次に単語の意味が書かれている。我々は、このようなデータのうち、語根からの現在語幹の生成については類・態に関する情報を用い、また現在語幹の自動生成に関する評価のために各種活用形に関する情報を用いることとした。

[6] 第 2 種活用においても第 1 種活用と同様に、それぞれの類の中には相当量の「例外」が含まれている。

[7] 語根が母音で終わる場合、"nu" は v/m で始まる語尾の前で "n" になることがある。また母音で始まる語尾の前では "nv" となる。語根が子音で終わる場合、母音で始まる語尾の前では "nu" は "nuv" となる。

[8] 語根によっては、子音で始まる人称語尾との間に "i" が挿入される。

[9] P.3.pl. の Present が "ati"、Imperfect が "ur" (さらに語根が guṇa 化する)、Imperative が "atu" となる。

[10] Parasmaipada(P) か Ātmanepada(A)。P,A の両方が可能なときは Ubhayapada(U)。

aṃh	v. 1.A; *(aṃhate, aṃhituṃ)* 1.to go
aṭh	v. 1.U; 1.to go
akṣ	v. 1.5.P; *(akṣati, akṣṇoti, ānakṣa, akṣiṣyati-akṣyati, ākṣīt, akṣitum-aṣṭum, akṣitvā-aṣṭvā, aṣṭa)* 1.to reach; 2.to pass through; 3.to accumulate
utpat	v. 1.P; 1.to fly or jump up; 2.to rebound (as a ball); 3.to rise; ...
pat	v. 1.P; *(patati, patita)* 1.to fall; 2.to fly; 3.to set (below the horizon); 4.to cast oneself at; 5.to fall (in a moral sense); ...

図 9: 高島 [8] の記述例 (動詞)

ところで、図 9 の中に "utpat" という見出し語が存在しているが、これは動詞語根 "pat" に接頭辞 "ud"[11]が接続したものである。このような単語が実際に文中で用いられる場合、単語の活用等は "pat" に基づいて行われ、その活用の結果生じた単語に "ud" が接頭辞として接続することになる。それゆえ、たとえば直説法過去形では過去を示す接頭辞 (augment) "a" が語幹の前に付与されるため "utpat" は "udapatat" (Impf.,P.,3,sg.) となる[12]。このような問題は、現段階ではとくに (3) 類の動詞語根の扱いにおいて問題となる。(3) 類では現在語幹生成の際に語根の先頭に重複辞が付与されるため、語根部分が確定できない状態では重複辞の作成、重複辞の接続の処理が不可能だからである。

そこで我々は、手作業により接頭辞付きの動詞の語根部分に区切りを入れる作業をおこなっている。作業の手順であるが、まず辞書から動詞の見出し語を取り出す。それらに対し、いくつかの経験則に基づき、動詞語根部分と接頭辞部分の切れ目と思われる箇所に自動的に区切りを入れていく。この自動的な区切り結果に対し、手作業により確認・訂正をおこなっていく、というものである。現在のところ、(3) 類動詞語根 101 個を中心として、全動詞の約 1/4 について、区切り箇所に関するチェックを終了している。

3.3 語根からの現在語幹の生成

前節で述べたような規則に従って、動詞語根から

[11] "ud-pat" が接続した結果 "utpat" となっている。
[12] Impf. に augment が付かないこともある。[10, p.115]

vṛ (5U)		
→	vṛṇ (5A):	弱, (+v,m)
→	vṛṇv (5A):	弱, (+母音)
→	vṛṇo (5A):	強
→	vṛṇu (5A):	弱

図 10: 語根から生成される動詞語幹

cl.	All	Ok		cl.	All	Ok	
(1)	1264	1202	(95%)	(2)	70	57	(81%)
(10)	712	601	(84%)	(3)	35	12	(34%)
(4)	149	113	(75%)	(5)	63	54	(85%)
(6)	182	154	(84%)	(7)	32	28	(87%)
(D.)	11	11	(100%)	(8)	30	18	(60%)
				(9)	104	82	(78%)
[1]	2318	2081	(89%)	[2]	334	251	(75%)

表 2: 自動生成された現在動詞語幹の精度

現在語幹を自動生成するシステムを用意し、高島 [8] から抽出した動詞語根のすべてに対し、動詞現在語幹の自動生成をおこなった。ここでは、たとえば語根が母音で終る (5) 類の弱語幹は "v" か "n" で始まる人称語尾の前では接尾辞は "n" になる、などの規則に対応するため、図 10 のような条件付きの語幹の生成をおこなう。

現在動詞語幹の自動生成がどの程度の精度で行われるかを確認するため、図 9 で示した例のうち、"akṣ" のように動詞活用形一覧が載せられている動詞語根を利用した。すなわち、動詞語根から自動的に生成された現在語幹に人称語尾を接続させたものが、辞書中に記述されている活用形、たとえば "akṣ" の例では "akṣati" あるいは "akṣṇoti" と一致しているかを調査するのである。この調査の結果を表 2 に示す。

表 2 からは、第 1 種活用動詞のほうが第 2 種活用動詞より文法規則どおりに現在語幹が生成される確率が高いことがわかる。ここで、文法規則どおり

cl.	All	[A]	[B]	[C]	[D]
(1)	62	14	35	12	1
(4)	36	2	26	0	8
(10)	111	0	100	9	2
(8)	12	0	12	0	0
(9)	22	2	12	5	3
[1]	237	19	185	21	12
[2]	80	16	46	5	13
All	317	35	231	26	25

表 3: 現在語幹生成の失敗の内訳

に語幹が生成されなかったと判断された事例について、その内容を整理したものを表 3 に示す。"[A]" は法則的な対処が困難と思われるもので、たとえば語根 "gam"(1) の現在語幹が "gaccha" であったり、"dṛś"(1) が "paśya" であったりするものである。"[B]" は一般的な文法規則からは導出できないが、文法書にない独自の法則を用意することによって対処可能と考えられるものである。"[C]" は辞書の記述が不十分な場合など、本節における評価の対象としては不適当と判断したものである。

このうち、対処が可能と思われた [B] の中では、現在語幹生成の際に起こる動詞語根中の母音の階梯の変化が、辻 [10] に記述されているものと異なっていた事例が 183 例あり [B] 全体の約 8 割を占めた。"śam"(10) が "śāmaya" となる[13] ような事例である。このような事例に対処するためには、動詞語根からの語幹生成の際に、語根中の母音の階梯をさまざまに変化させたものを同時に生成させる方法が考えられる。たとえば語根 "nī"(1) からは語幹 "**naya**" が生成されるが、それ以外にも "ī" と同じ階梯のグループに属する "y","iy","āy" を用いた "**nya**" "**niya**" "**nāya**" などの候補も同時に現在語幹の候補として生成するのである。また [B] の中には、動詞語幹生成の際に鼻音を追加あるいは削除する事例が 34 例あった。"olaj"(1) から "olañja" という語幹が生成されるような事例である。これも母音の階梯と同様に、語幹生成の際に複数の語幹候補を生成するようにシステムを直すことは容易である。しかし、無限定に語幹候補を複数生成してしまう方法は、形態素解析における曖昧性を生じさせる原因になる可能性が高い。それゆえ、複数の語幹候補の生成については、今後慎重に検討を重ねる必要がある。

3.4 語幹と語尾の接続

実際に文中に登場する動詞は、動詞語幹に人称語尾が接続したものである。この接続の際に問題となるのが、動詞語幹末尾の文字と人称語尾先頭の文字との連声である。

現在動詞語幹のうち第 1 種活用に属するものは、いずれも動詞語幹末尾文字が "a" であるため、連声の問題はほとんど生じない。同様に (5),(8),(9) 類も、

[13]「am で終る語根は一般に a を延長しない」[10, p.174]

word	stem		ending
-tthās	-d	+	thās
-ddham	-dh	+	tam
-ddhām	-dh	+	tām
-ḍha	-h	+	tha
-ḍhās	-h	+	thās
-gdhās	-h	+	thās
-ddhvam	-th	+	dhvam

図 11: 単語から語幹・語尾組への対応表

語根部分の後に接尾辞が付けられているため、類単位での連声の処理が可能である。しかし (2),(3),(7) 類については、語幹末尾の文字が直接人称語尾と接続するため、連声の扱いが複雑になる。たとえば動詞 "leḍhi" は語幹 "leh"(2) から生成された現在語幹に人称語尾が接続した "leh-ti" であるが、そのように解析されるためには連声による音韻変化に解析システムが対応しておく必要がある。

我々は、あらかじめ音韻変化に関する表を作成しておき、その表を用いて解析をおこなうこととした。まず、(2),(3),(7) 類動詞語根すべての末尾文字を調べた。その結果、末尾文字としては "ā", "j", "d", "u", "s" など 26 通りの可能性があることが判明した。これら 26 種類の語幹末尾子音と、すでに図 8 に示した人称語尾の先頭文字 "s", "t(th)", "dh", "m", "v", "y" および母音との間で発生し得る連声のパターンをあらかじめ展開してしまうのである。そのうえで図 11 に示すように、ある特定の単語末尾文字列から、語幹末尾・人称語尾の組の情報を得ることができる表を作成する。これによって、連声への対処をおこなうこととした。

4 おわりに

本稿で我々は、古典サンスクリット語動詞のうち、現在組織を対象とした単語の形態素解析を実現すべく、以下の点について述べてきた。

- 動詞語根からの現在語幹の自動生成とその評価
- 連声を考慮に入れた、動詞語幹と人称語尾との組合せ表の作成

このうち前者においては、高島 [8] に活用についての記述がある動詞見出し語 2652 個のうち、現在動詞語幹との対応表を手作業によって用意する以外

に方法がないものが 35 個、また辻 [10] では述べられていない法則を用意することにより対処できそうなものが 231 個、合計すると 266 個 (10.1%) が例外的な処理を必要とするものであることが明らかとなった。この割合を高島 [8] に記載されている動詞全体に当てはめると、約 500 語程度の動詞について、今後さらに何らかの対処をおこなう必要があることになる。このうち、新たな規則の追加によって対処できるものについては、検討のうえで何らかの対処をおこなっていきたい。

また後者においては、あらかじめ総当たり的に連声も含めた人称語尾のパターンを展開して表にしてしまうことにより、連声に関する処理を単純化している。

今後の課題であるが、現在のところ、我々が構築した解析システムに対して定量的な評価を下すための実験データが用意できていないため、システムの評価ができない状況である。それゆえ、我々が構築したシステムの評価をおこなうことが目下の最大の課題である。

また本稿では言及しなかった、現在組織以外の動詞活用組織に対する対処についても、今後は考察の対象に加えていきたいと考えている。実用に堪える古典サンスクリット語の形態素解析システム構築のためには、単に単語の曲用や活用を扱うだけでは十分ではなく、単語間の連声や複雑な複合語も処理できる枠組を用意していく必要があるが、それはまだ遠い先の話になると思われる。

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、また高島先生との交渉の窓口になってくださった永崎研宣さんに感謝いたします。

インド仏教学の観点における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東京大学東洋文化研究所の鈴木隆泰さん、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がたに感謝いたします。

古典サンスクリット語の動詞規則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] V.S. Apte. *Practical Sanskrit-English Dictionary*. Poona, revised & enlarged edition, 1957 (Repr. 1978 in Kyoto).

[2] The Asian Classics Input Project. WWW. (Jul. 13, 1999) URL: <http://www.asianclassics.org/>.

[3] Chinese Buddhist Electronic Text Association (CBETA). WWW. (Nov. 12, 1999) URL: <http://ccbs.ntu.edu.tw/cbeta/>.

[4] 上村勝彦. カウティリヤ 実利論 (2 vols). 岩波文庫. 岩波書店, 1984.

[5] Sri Lanka Tripitaka Project. Journal of Buddhist Ethics – Pali Canon Online. WWW. (Jul. 14, 1999) URL: <http://jbe.la.psu.edu/ibric.html>.

[6] 鈴木隆泰. Tibetan-Sanskrit 構文対照電子辞書プロジェクト eDic. WWW. (Oct. 10, 1999) URL: <http://www.info.ioc.u-tokyo.ac.jp/suzuki/edic/>.

[7] 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.

[8] 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/>.

[9] The Association for Computerization of Buddhist Texts (ACBUT). SAT – machine-readable text-database of the taisho tripitaka (the taisho shinshu daizokyo). WWW, 1998. (Nov. 12, 1999) URL: <http://www.l.u-tokyo. ac.jp/~sat/>.

[10] 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.

[11] M. Yano. FTP. (Nov.12, 1999) URL: <ftp://ccftp.kyoto-su.ac.jp/pub/doc/sanskrit/>.