2004－CH－63　(5)
2004／7／30

語幹情報に注目したサンスクリット文献閲覧システムの試作

相場徹 †　　生出恭治 ‡

概要: 古典サンスクリット語 (Skt.) の電子テキストを自在に使いこなすことは難しい。

本稿では Skt. の電子テキストの利便性向上について考察する。我々は、語幹や活用などの文法情報を電子テキスト中の各単語に付加することによって、電子テキストの利便性が向上すると考える。そこで、文法情報を自動的に電子テキスト中の各単語に付加することを試みた。

また我々は Skt. の電子テキストの閲覧システムを試作した。このシステムは、電子テキスト中の各語幹ごとの索引を表示できる。我々のシステムはまだ十分な利便性を提供できるには至っていないが、今後、電子テキストに付与する文法情報の信頼性を向上させることによりさらに便利なツールになると考えている。

Browsing System for Sanskrit E-Texts: Using the Information of Word Stems

Tooru AIBA †　　Kyoji OIDE ‡

Abstract: Currently we have many more e-texts in classical Sanskrit, but we still cannot easily analyze them by using software tools now in existence.

In this paper we try to seek improvements of the usability of e-texts in Sanskrit. We propose the grammatical information of each word such as stems and inflection should improve the usability of Sanskrit e-texts. So we have tried to automatically append such information to each word in an e-text.

Then we have tried to implement a browsing system for Sanskrit e-texts, which can show the indexes of each stem. Our system has not yet been usable enough for Sanskrit researchers, but it will work as a helpful tool when we improve the reliability of the grammatical information of each word.

† 東北大学 大学教育研究センター Research Center for Higher Education, Tohoku University.
‡ 東北大学 Tohoku University.

1 はじめに

昨今は古典文献の電子化がかなり急速に進んでいる。しかし、膨大な分量となってきた古典サンスクリット語 (Skt.) の電子テキストを本格的に活用して行われた研究成果を目にすることは少ない。

そこでまず古典学、とくに本稿で中心的に扱う Skt. の文献を扱う学問分野における「研究」の方法を確認し、その方法において望ましい電子テキストについての検討を行いたい。

1.1 古典学の研究対象・方法

古典文献、とくに古典インド文献を扱う際に問題となりやすいのが、文献の成立過程である。たとえばヴェーダ文献や仏教経典など聖典と呼ばれる文献の多くは現存する文献がそのまま文献成立当初の姿を伝えるものではなく、伝承の過程において、さまざまな地域や時代に付加増広が行われた結果、現存するテキストができあがったと考えられている。すなわち Skt. 文献の特徴の一つに次のような点があげられる:

- 古い時代からの伝承と、後代における付加増広部分とが混在。後代の付加増広部分は、伝承の途中で徐々に追加されたもので、さまざまな時代の、さまざまな要素が混在
- 各文献の成立 (付加増広) の年代および背景を知る手がかりとなるのは、記述内容 (物語の展開など) や言語的特徴など

このような特徴をもつ Skt. 文献を対象とした研究の方法について、以下に 2 つ例を紹介する。

Kasamatsu [6] Skt. 以前の古代インド語における「頭」(síras-/śīrṣáṇ-) という語に注目し、この単語をヴェーダなどのバラモン教関連の諸文献から取り出した。こうして取り出した単語には、たとえば sing. Loc. の語形だけを取ってみても、RV[1] には “śīrṣán”, “śīrṣáṇi”、YS には “śīrṣán” および例外的に “śīrṣaṇi”、AV には “śīrṣe” “śīrṣáṇi”、ŚBK には “śírasi” など、さまざまな語形がある。これら各語形を比較し、類似関係を整理することにより、たとえば RV と YS はいずれも同じ “śīrṣán” を用いているため、両者の言語層は比較的近いものであるが、一方で両者は AV との言語的関連が少ないことから AV とは異なる言語層に属すると推定される。また古い時代に成立した文献のうち ŚBK は一般的な Skt. と同じ活用形を用いており言語的に特異である。こういった各文献における語法が、用例を含めた精密な読解に基づいて報告されている。

岡野 [7] Skt. で書かれた大乗仏教経典のひとつ *Lalitavistara* (Lv) の成立について論じている。まず現存する漢訳のテキスト『普曜経』と Lv の内容を詳細に対比させ、現存する Lv に含まれる記述のうち、両方のテキストに共通して含まれる文を比較的古い時代に作成された「原形」、それ以外を「付加部分」に分類した。岡野はさらに、この「原形」の各部に見られる記述の不整合に注目し、そこから「原形」に先立つ「原初形態」の存在を推定する。そして「原形」では、Lv において重要とされる章ほど世俗的な説話や仏伝等でよく用いられる《前 Campū 様式》と呼ばれる表現形式で書かれていること、また《長行重頌》と呼ばれる、聖典でよく用いられる形式で書かれた個所の表現の不自然さから、 Lv の「原初形態」は《前 Campū 様式》のみで構成されていたと推定する。このことから、最初から意識的に聖典たるべく作られた『法華経』などとは異なり、Lv は最初は雑多な過去の韻文による民間伝承の雪だるま式な集積であったと結論付けられている。

1.2 電子テキスト有効活用への対処

前節の Kasamatsu [6] および岡野 [7] の論述の中で、我々は特に以下の点に注目する。

- 特定の単語 (語幹) に関する情報の収集
- 同一文献あるいは他文献との、類似部分 (parallel) に関する情報の収集

このうち前者については、単語の出現および活用形などの調査は、基本的には文字列の単純なパターンマッチを調査するだけでよいので電子テキストの有効活用が比較的実現しやすいと考えられる。他方、後者については、当該段落の記述内容、また場合によっては前後の段落をも含んだ文脈の理解など、かなり高度で複雑な知識処理が要求されるため、現状

[1] 本段落の略号: RV: *Ṛgveda*; YS: *Yajurveda Saṃhitā*; AV: *Atharvaveda*; ŚBK: *Śatapathabrāhmaṇakāṇva*.

ātman
ātmā
a1) bandhur**ātm*ā***tmanastasya (BhG 6-6a)
a2) tv**ātm*ai***va (BhG 7-18b)
a3) tath***ātmā*** (BhG 13-32d)
ātmanas
b1) hy**ātmano** (BhG 6-5c)
b2) bandhurātm***ātmanas***tasya (BhG 6-6a)
b3) yogam**ātmanaḥ** (BhG 6-19d)

図 1: 文中における “ātman”

では電子テキストの利便性を活かすことは困難であろう。そこで本稿では、前者に焦点を当て、この目的に沿って電子テキストについて考えていきたい。

先に、前者については文字列の単純なパターンマッチを調査するだけでよいと述べたが、Skt. 文献を扱う際にはそれさえも様々な困難を含んでいる。この困難の要因となるのは、1) 単語が格変化・活用すること、2) sandhi (連声) と呼ばれる音韻規則により単語の先頭・末端文字が文字変化すること、である。

図 1 に、名詞 “ātman” が実際に *Bhagavadgītā* (BhG) のテキスト中にどのように出現しているかを示す。名詞 “ātman” は “ātmā” “ātmanas” 等の格変化をした語形が、さらに文中に出現する際に文字変化を起こす。たとえば 図 1 の “ātmanas” の例では、次単語の先頭文字に影響され語末の “as” が b1 では “o”、 b2 では不変、 b3 では “aḥ” と、それぞれ別の文字に変化している。また変化の結果、単語間の境界部分の文字が融合することもある。たとえば “ātmā” の a1 の例では “ātmā” の語末の “ā” が後続語 “ātmanas” の先頭の “ā” と融合して “ātmātmanas” となっている。また a2 の “ātmaiva” は、“ātmā” の語末の “ā” と後続語 “eva” の先頭文字 “e” とが融合して “ai” となった結果である。

これらの事例から、Skt. の電子テキストを扱う際には、以下のような対処が必要となる: 1) 融合した複数単語の切り分け、2) 単語境界部分にあった本来の文字の推定、3) 単語の格変化・語形変化への対処、である。

1.3 単語情報付加と閲覧システム

このような Skt. の困難を乗り越えるため Kyoto archives of Sanskrit texts の *Mahābhārata* [11] は、単語をすべて手作業で切り分けた状態を電子テキスト化している。この方法で構築されたテキストは、データ検索などの用途では非常に利便性が高いが、底本となっている紙媒体のテキストの内容が完全には復元され得ない点、および、特定の単語語幹・語根に関連した情報の収集が難しい点に不満が残る。

それを解消するためには、電子テキストに何らかの方法で単語情報や単語活用情報を埋め込む方法が考えられる。たとえば図 1 a2 の “tvātmaiva” は “tu” “ātmā” “eva” という 3 つの単語によって構成されるが、ここに、その文字列が 3 単語によって構成されるという情報、および “ātmā” は “ātman” という辞書見出しの sing. Nom. の活用形であるという情報を、電子テキストに付け加える方法である。

しかし、実際に単語を手作業で区切り、さらに単語の活用情報も加える作業は非常に手間と時間がかかり、現実的とは言えない。とくに Skt. は文法の複雑さ、また言語知識を持つ作業者が希少である点などにより、作業は一層困難となる。そこで本稿では、まず、なるべく人手によらない Skt. の電子テキストへの単語・文法情報の追加を試みる。

ところで、電子テキストに埋め込む情報として、ここまで我々は単語の語幹情報や活用情報をあげてきた。しかし、これ以外にも電子テキストの利便性を上げる情報が何かあるのではないか。このように考えた我々は、現実にどのような情報があると便利かを模索するための道具として、単語の語幹・活用情報を付加した電子テキストを用いた閲覧システムを試作し、そのシステムを実際に使用してみることとした。本稿では、我々が試作した閲覧システムについても述べる。

2 電子テキストへの情報追加

前節でも述べたとおり、Skt. の電子テキストを扱う際には、sandhi 等によって生じた文字変化、また分かち書きされない単語をどのように区切るか、といった問題がある。

Sandhi Table			Reverse Table		
t + ch	→	c+ch	c+ch	←	t + ch /
d + ch	→	c+ch			d + ch /
d + ś	→	c+ch			d + ś
d + j	→	j+j	j+j	←	d + j
d + ph	→	t+ph	t+ph	←	d + bh
d + v	→	d+v	d+v	←	d + v

図 2: 音韻変化 (sandhi) 表

2.1 sandhi への対処

単語の切り分け sandhi によって結合した単語の切り分けの自動化は現状では実現できていない。自動的に切り分けを行うには、解析の際に生じる膨大な量の解析結果候補のうち、それらしい候補以外を抹消する処理が必要となる。解析候補の絞り込みの際によく用いられるのが品詞の前後関係や各語幹・活用形などの出現確率などの情報であるが、Skt. においては、まだそのような言語的傾向に関する計量的な研究成果が得られていない。それゆえ結合した文字の切り分けの自動化は今後の課題とし、本稿においては手作業で行わざるを得ないと判断した。

この段階で、たとえば “tad śrutvā” が結合した “tacchrutvā” に対し、手作業で “tac_chrutvā” のような形で区切り情報の追加を行う。

文字変化への対処 sandhi によって生じる文字変化はパターン化されており、機械的な処理が比較的容易ではないかと考えられる。そこで我々は、辻 [12] の sandhi に関する記述に基づき、図 2 のような対応表を作成することによって、sandhi による文字変化への自動的な対応を試みることとした。

図 2 の左側の「Sandhi Table」が、辻に基づいて作成した表の一部である。この表は sandhi が起こる前から後への文字の変化を示すものであるが、この「前から後へ」の情報は、本稿において我々が求める「sandhi の適用後から適用前の状況を求める」処理とは方向が逆となる。そこで「Sandhi Table」の情報を利用して、「後から前へ」の情報を得るため、図 2 の「Reverse Table」で示したような逆表を作成した。この逆表は、電子テキスト中の (手作業によって切り分けられた) 単語から、sandhi 適応前の語形候補群を生成するときに利用できる。

ここで注意すべきは、sandhi の適用前と適用後とは 1 対 1 対応ではない点である。たとえば “tac chrutvā” の sandhi 適応前の状態は、図 2 の逆表の最初の規則に従うと、正解である “tad śrutvā” 以外にも、“tat chrutvā” “tad chrutvā” の可能性を持つ。これら正解以外の候補を排除するため、我々は語彙情報を用いる。それゆえ、次節に示すような単語の語形解析を上記 3 候補に対してそれぞれ行い、その結果、きちんと解析結果が得られた候補のみを適切なものとして残す、という対処を取る。

2.2 単語活用情報付与の自動化

本稿において各単語に付与しようとする文法情報とは、主に以下のものである。

- 辞書見出しとの対応情報
- 活用形がある品詞のときは、活用情報

本稿では Apte [5] に基づいた Takashima [9, 10] を基準とし、この辞書に掲載されている見出し語との対応付けを試みることとした。Skt. における単語の活用体系は、おもに名詞・形容詞型活用と、動詞型活用とに分けられる。我々は、この両方について自動的な語形解析を試みている最中である。以下にその概略を述べる。

2.2.1 名詞・形容詞型活用

Skt. 文中に出現する名詞・形容詞の各単語は、語幹と格語尾という 2 つの要素より成っている。

語幹 Skt. の名詞には、複数の語幹の種類が定義される。語幹の種類は、名詞の辞書見出し末尾の文字 (列) で判別できる。たとえば “aśva” という辞書見出しは末尾文字 “a” から「a-語幹名詞」に、“ātman” は末尾の “man” から「man-語幹名詞」に分類される。我々は Takashima から名詞・形容詞の単語見出しを抽出し、語幹文字列の一覧を用意した。抽出できた語幹の数は、名詞が約 27000 項目、形容詞が約 11000 項目となった。

格語尾 図 3 に示すように、Skt. では語幹の種類ごとに格語尾が異なることが多い。それゆえ、辻を参照して語幹の種類ごとの格語尾表を用意した。

我々は、このような語幹および活用語尾に関する情報 (文法セット) を用いて単語語形解析を行う

	a-語幹			i-語幹	
	m. sg.	n. sg.	m. pl	m. sg.	m. pl.
N	aḥ	am	āḥ	iḥ	ayaḥ
Ac	am	am	ān	im	īn
I	ena	ena	aiḥ	inā	ibhiḥ
L	e	e	esu	au	isu
V	a	a	āḥ	e	ayaḥ

図 3: 名詞・形容詞の格語尾の例

簡単なパーザを構築した。このパーザは、たとえば “aśvena” という単語を、語幹辞書から得られる “aśva” という a-語幹名詞 (m.) と、格語尾表にある “ena” という a-語幹名詞格語尾 (m. sg. I.) との組み合わせだという解析結果を出力する。

2.2.2 動詞型活用

動詞も名詞と同様に、基本的には語幹部分と活用語尾部分とによって構成されるが、動詞の辞書見出しが語幹 (stem) ではなく、動詞語根 (root) であるため、動詞の処理は名詞よりも難しい。Skt. では、語根から時制等に応じて語幹が生成され、その語幹と語尾とが接続して活用形となる。たとえば “gacchati” は動詞現在語幹 “gaccha” と人称語尾 “ati” が接続した結果であるが、この単語に対応する辞書見出しは動詞語根 “√gam” であるため、“gacchati” と “√gam” の対応付けが求められるのである。

このような動詞活用形の語形解析を可能とするため、我々はあらかじめ辞書中の各動詞語根から各種動詞語幹を自動生成しておき、この各種動詞語幹と活用語尾とを組み合わせて解析するのがよいと考えた。我々は Takashima にある各動詞語根からの各種動詞語幹の自動作成を試みている[2]。我々は、こうして構築した語幹情報、および手作業によって用意した活用語尾一覧を用い、動詞活用形を解析する簡単なパーザを構築し利用している。

2.2.3 その他

接続詞などの不変化詞については、まず辞書から不変化詞の見出しを抽出しておき、それと解析対象文字列とが一致するかどうかを調査する。

また指示代名詞・人称代名詞は名詞と同様に格変化を行うが、対象単語数が少ないことから、すべての格変化形を語彙としてデータに追加し、不変化詞と同じように、語彙と文中単語の完全一致を調べるだけで解析できるようにした。

[2]各種動詞語幹の構築については、[3, 4] 等で述べている。

品詞	語数	解析可	
動詞	70	65	(92.8%)
動詞 (分詞)	40	32	(80.0%)
名詞	221	157	(71.0%)
形容詞	56	44	(78.5%)
計	387	298	(77.0%)

図 4: BhG 第 1 章の語形解析精度

2.3 電子テキストへの単語情報の付与

我々は、前節で述べたようなパーザを用意し、BhG の電子テキスト (約 10000 単語) を対象とした文法情報の付与を行った。自動的に付与された文法情報の精度確認のため、我々は菅沼 [8] を利用した。菅沼の BhG 第 1 章における各単語の活用形を解説する章での記述とパーザの解析結果とを比較し (不変化詞、代名詞等は除外した)、パーザが出力した複数の解析結果のうち、どれか一つでも菅沼 [8] の記述と一致したものがあると「解析可」とした。こうして行った数え上げの結果を、図 4 に示す。

一般的に、複数の解析結果が得られたときは、各解析結果がどの程度尤もらしいかを数値的に表現し、その尤もらしさの数値が最も大きいもののみをパーザの解析結果として扱うことが多い。しかし我々の現状では、それぞれの解析結果の尤もらしさを測定する基準に関する手がかりが全くない状態にある。そこで、ここでは「その単語の語形を解析できる文法セットが用意されているかどうか」を見るという観点から、「解析可」という基準での文法セットの評価となっている。尤もらしさの判定基準については、今後、計量文献学的な方法による文献へのアプローチを強めることにより、何らかの手がかりを得ていく必要があろう。

さて図 4 からは、名詞・形容詞の解析精度が比較的低いことが言える。この理由としては、BhG 第 1 章には人名がかなり頻繁に登場しており、これが Takashima の辞書見出しと一致しにくいため解析に失敗していること、また名詞・形容詞に対する我々の現段階での文法セット構築が動詞ほど進んでいな

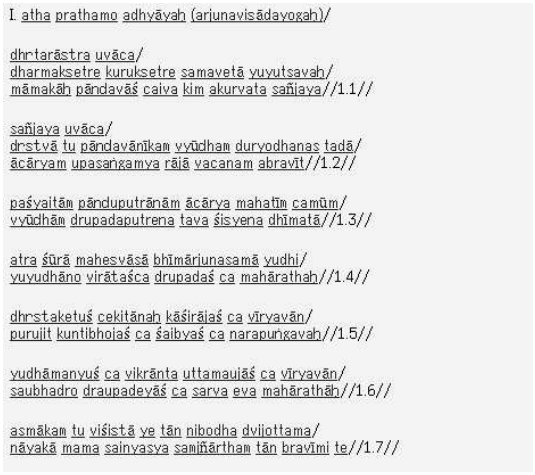

図 5: テキスト閲覧画面

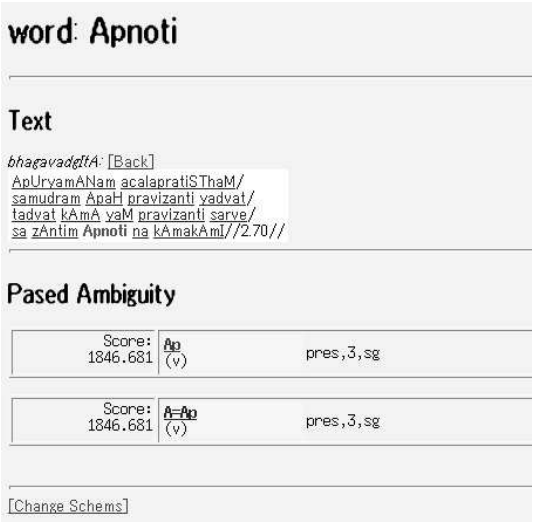

図 6: 単語情報画面

いこと等が主な原因として挙げられる。

3 閲覧システムの試作

我々は、ギリシア語やラテン語などのテキストを対象とした電子テキストの閲覧システム、The Perseus Digital Library[1] を参考にしながら、電子テキスト閲覧システムの試作を行っている。

現在我々が試作したシステムは WWW の CGI プログラムとして動作するものであり、「テキスト閲覧」「単語情報」「単語語幹出典一覧」という 3 種類の画面で、それぞれ異なる情報を提示する。

3.1 テキスト閲覧

テキストと付加情報 同じ画面上にさまざまな情報が混在するのは使い勝手が良くないと判断し、図 5 に示すように、付加情報は一切表示しないようにした。その代わりに、各単語につけられたハイパーリンクをクリックすれば、各単語に付与された文法情報が表示される。

なお 2.1 節で問題とした、sandhi によって融合した文字の扱いは以下のようにした。たとえば Skt. 文中の “caiva” は “ca” “eva” という 2 単語によって構成されるという情報を電子テキストに加えていたが、このような情報を活かすため、テキスト閲覧画面では “caiva” と表示しておき、単語情報のアクセスの際に “cai” 部分をクリックしたら “ca”、“va” 部分をクリックしたら “eva” に関する単語情報が示されるようにした。

原典の表記方法 古典語をコンピュータで扱う際には、表記文字・転写方式が問題となることが多い。Skt. はダイアクリティカルマーク付きアルファベットでの表記が基準となるが、その文字を表記するために、以下のような方法がある [2]。

- 置換表記する方法 (KH, TS, ITRANS etc.)
- 特殊な文字コードを割り当てる方法 (Normyn, Pali96, Unicode (UTF-8) etc.)

前者は、たとえば “ātman” を “Atman”(KH) “aatman” (Kamimura) のように、特殊文字を他の文字で表記する方法である。後者は、一般に ASCII 等で用いられない文字コード領域に “ā” などの文字コードを割り当てて使う方法である。

このうち、我々は「置換表記する方法」の主要なものと、「特殊な文字コードを割り当てる方法」のうち Unicode を用いたテキスト表示とも利用できるようにし、どの方法で画面表示するかをユーザ自身が選択できるようにした。

3.2 単語情報

図 5 のテキスト閲覧画面でどれか単語をクリックすると、図 6 のような単語に関する付加情報、すなわち単語の語幹および活用に関する情報を示すページに変わる。

ここでは、前節で述べた語形解析パーザの出力結果を、辞書見出し (語幹・語根) および品詞ごとに列挙して表示される。

単語語形解析システムの問題点　単語情報画面を実際に作成して感じた問題点として、我々の語形解析パーザが出力する解析結果の数が多いことが挙げられる。図 6 の例では、単語 “āpnoti” は動詞語根 “$\sqrt{\text{ā}}$p” か “ā$\sqrt{\text{ā}}$p” のいずれかから生成された語形であるという、2 つの可能性が示されている。このうち “ā$\sqrt{\text{ā}}$p” から作られる動詞語はおそらく実際の Skt. 文中には出現しないと思われるが、我々が用意した語形解析システムがこのような候補を排除できず、解析結果候補として残ってしまったものである。実際に試作した閲覧システムを試用した結果、このように大量の誤った解析候補が列挙されてしまうことが時々あり、「単語情報」画面が非常に使いにくく感じることが多いと実感した。我々が構築した語形解析パーザは、我々が用意した語幹一覧、語尾一覧、および図 2 に示したような sandhi 表という 3 つの文法セットを機械的に組み合わせて語形解析を行うが、これらの情報に関する文法書の記述、とくに sandhi が発生する条件に関する記述にやや厳密性が欠けているため、3 つの文法セットを組み合わせて計算機処理を行うと、組み合わせ方によっては非現実的な解析結果が生じてしまう可能性を排除できない。この問題に対する対処については、我々のパーザが生み出した非現実的な解析結果の原因を探り、語幹一覧・語尾一覧・sandhi 表を訂正していくという地道な対処が必要となる。

3.3　単語語幹の出典一覧

ここまで紹介してきた「テキスト閲覧」「単語情報」の各画面は、我々が用意した電子テキストが持つ情報を単純に表示するものであったが、語幹情報付きテキストの利点を活かし、語幹レベルでの出典情報一覧を表示できるようにしたのがこの「出典一覧」画面である。図 6 にある 2 つの解析結果候補のうち “āp” 部分をクリックすると、図 7 に示したように、“āp” という解析結果が得られた単語を含む文を BhG の登場順に一覧表示する。

ここでの問題は、我々の単語語形パーザの精度の低さが与える影響である。図 4 に示したとおり、文

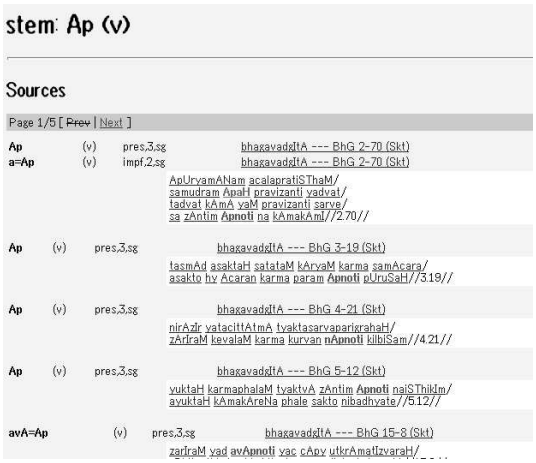

図 7: 語幹の出典一覧画面

中単語全体の約 1/4 は現段階での我々のパーザでは解析できない。解析に失敗した単語については、現在のところ、この語幹レベルでの出典情報一覧の対象には含まれないため、文中単語の約 1/4 はここでは除外されていることになる。それゆえ、ここでの利便性向上のためには、我々の語形解析パーザにおいて、図 4 で述べた「解析可」の数値を上げることが当面の課題となろう。

4　評価

我々が試作した文献閲覧システムを評価するため、1.1 節で紹介した Kasamatsu および岡野の研究方法に対し、現段階での我々の閲覧システムがどのように貢献し得るか、またどのような改良を行うことが望ましいか、などの点について考えてみる。

Kasamatsu の研究は単語の語形の変遷をヒントに、各文献におけるいわば「言葉遣い」の傾向を明らかにするものであった。どれか一つの単語に注目し、その単語に関連する語形を収集して分析するという方法そのものは、本システムが想定する研究方法と合致するものである。それゆえ今後、我々が用いるパーザの解析精度の向上、閲覧できるテキストの分量増加などによる「単語語幹の出典一覧」画面の資料的価値の向上によって、我々の閲覧システムは実用的なものに発展することが期待される。実際にシステムを試用した筆者の個人的な印象では、「単語語幹の出典一覧」の画面において、現状では当該

語幹を含む文を登場順に出力するだけのものであるが、たとえば活用形ごとに並べ変える、あるいは、活用形ごとの何らかの傾向を分析した情報などが加えられると良いのではないかと感じた。今後どのような情報を電子テキストに追加していくと有用か、また閲覧システムにおいてどのように情報を表示すると便利であるか等については、今後実際にさまざまな試行錯誤を行っていく必要がある。

逆に問題となるのは、単語語形パーザが前提とする部分である。Kasamatsu は変則的な単語活用に焦点を当て、各文献の言語的性格を読み解こうとするものであった。しかし我々のパーザは、標準的な Skt. の辞書および文法書に従って構築した文法セットに基づいて語形解析を行うため、現行の枠組みのままでは Kasamatsu が重視する単語を扱うことができない。よって、今後は Skt. などの古典言語には比較的多い、変則的な活用の語形にいかにして対処するかが課題となろう。

岡野は Lv と漢文の文献との記述内容の比較を行い、両者に共通して述べられている部分を「原形」と見なした。このような研究の方法に対しては、電子テキスト中の各文あるいは段落ごとに、同一文献あるいは他の文献における類似部分についての情報が付けられていると非常に役立つと思われる。しかし現状では、このような種類の情報の付加を手作業によらない形で実現することはかなり難しい。この点については、今後の課題としたい。

5 おわりに

本稿では、古典学の方法に留意しながら、Skt. の電子テキストに対する自動的な語幹・活用情報の追加、また、追加された語幹・活用情報を用いた閲覧システムの試作について述べた。

我々が試作した閲覧システムは、Skt. 文献を扱う研究のうち一部のものについては、

- 語形解析パーザの精度向上、古典文献には多い単語の変則活用への対処
- 閲覧可能な電子テキストの増強

これらの課題を解決していくことによって、研究者にとって実用的なものとなることが期待できる。

ただし、後者の課題については、膨大な分量のテキストが扱えるようにするためには、現状では手作業によってしか対処できない単語の切り分け作業がボトルネックとなろう。この問題を解消するためには、単語の切り分け作業の自動化が望ましい。これは今後の課題である。

我々が試作した閲覧システムは以下の URL からアクセス可能である:

`http://texa.human.is.tohoku.ac.jp/aiba/demo/`

我々は電子テキストに求められるであろう情報の追加を行い、その情報を利用した閲覧システムを実際に試用したが、これにより以後取り組むべき課題が次々と明らかになった。今後、これら課題を順に解決していきながら電子テキストおよび文献閲覧システムの有用性を向上させていきたい。

参考文献

[1] The Perseus Digital Library. WWW. (Nov. 14, 2003) URL: <http://www.perseus.tufts.edu/>.

[2] 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/>.

[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] 相場徹, 生出恭治. 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. 情報処理学会研究報告, Vol. 2002-CH-53-5, pp. 33–40, 2002.

[5] V.S. Apte. *Practical Sanskrit-English Dictionary*. Poona, revised & enlarged edition, 1957 (Repr. 1978 in Kyoto).

[6] S. Kasamatsu. On the inflection of OInd. *śíras-/śīrṣáṇ-*, n. 'head'. *Journal of Indian and Buddhist Studies*, Vol. 52, No. 2, pp. 1–5(L), 2004.

[7] 岡野潔. Lalitavistara の原初形態について. 印度学仏教学研究, Vol. 40, No. 1, pp. 70–74(L), 1991.

[8] 菅沼晃. サンスクリット講読. 1986.

[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/~tjun/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・パースカラーオ), pp. 73–105, 2000.

[11] M. Tokunaga. The digitalized text of *Mahābhārata*. WWW. (Jun. 23, 2004) URL: <http://www.kyoto-su.ac.jp/~yanom/sanskrit/mahabharata/>.

[12] 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.