東京大学東洋文化研究所附属東洋学研究情報センター報

明日の東洋学

October 2002 no.8

Research and Information Center for Asian Studies（RICAS）
Institute of Oriental Culture, University of Tokyo

クメール文字コードをめぐる諸問題

原田至郎

なぜぼくのWindowsでミニチュア・アリフが打てないのか

保坂修司

仏教学におけるN-Gramの活用

石井公成

本研究所所蔵のアラビア語写本のひとつ、ガザーリー（西暦1111年没）の『宗教諸学の再生』の一部である。ガザーリーはスンニー・イスラームにおいてイスラーム思想の伝統的な形態を確立した思想家と考えられるが、本書は彼の主著とみなされる著作である。ここに写っているのは、最後の編である第4編、魂の救済について述べている部分のなかで、人間の誠実さについての議論の終わりと人間の魂を注視し監視することを論じる章の始めの部分である。Cairo 1968年版ではVol. 4, pp. 488f.に対応する。ダイバー写本Ⅰ、No. 289より。（解説：鎌田繁　西アジア研究部門教授）

仏教学におけるN-gramの活用

石井公成

1. N-gramとの出会い

この数年は、内外の仲間たちの活動のおかげで、漢文仏教文献は、かなり電子データが揃ってきた。そうなると、人間というのは欲深いもので、異本の電子テキストも欲しくなってくる。また、異本や異訳などを簡単に比較してくれるシステムが欲しくなってくる。とはいえ、電子テキストを一字一句に至るまで厳密に校訂し、すべての文章に文法情報を加え、情報学の最先端の成果を活用して分析を試みるといったことは、筆者のように、ものぐさで理系の学問に弱い文系研究者には、とうてい無理なことである。ある程度の入力ミスを含む電子データでも使いものになる手法、それも簡単な操作で処理できるような方法でないと困るのである。その代わり、多くは望まない。コンピュータによって完璧に分析しようとか証明しようといった気持ちは、もともとないのだから、何かしらヒントが得られれば、それで十分である。

そうした横着かつ欲深な思いが強くなりつつあった際に出会ったのが、知友の近藤泰弘氏・近藤みゆき氏が国語学・国文学の世界に導入し、目からうろこが落ちるような成果をあげてみせたN-gramの技法であった。この技法は、近藤夫妻のようなコンピュータ利用の達人が慎重に用いれば、厳密な研究成果が得られる一方、筆者のような素人が雑に使っても「それなりの」結果を生み出してくれるものであったのだから、筆者が大喜びしたのは当然であろう。N-gramは、両氏を含む漢字文献情報処理研究会の仲間たちの間でブームとなり、あっという間に便利なツールが次々作成され、また様々な活用法が提案・報告されるようになった。そして、ついに同研究会の雑誌である『漢字文献情報処理研究』第2号（2001年10月）では、「N-gramが開く世界―確率・統計的手法による新しいテキスト分析」と題する特集を組むにまで至ったのである。

2. N-gramによる処理

N-gramとは、情報理論の祖であるクロード・シャノンが開発した確率・統計的自然言語処理の方法であり、テキストを任意の大きさの単位で自動的に分割したのち、特定の文字の組み合わせがどれだけ登場するかを計算し、その結果から様々な情報を引き出そうとするものである。

たとえば、「東京大学東洋文化研究所附属東洋学研究情報センター」という文字列を2字単位で分割してゆくと、次のようになる。

東京
京大
大学
学東
東洋
洋文
……

1字から6字までという指定で切ってゆくと、こんな具合になる。

東
東京
東京大
東京大学
東京大学東
東京大学東洋
京
京大
京大学
京大学東
……

いずれの方式を用いるにせよ、こうして分割していったうえで、特定の組み合わせが何回出てくるかを自動的に計算させるのである。

3. NGSMによる処理

上のような形で分割して出現回数をカウントすれば、その結果を他の文献について同様のやり方で処理した結果と比較することが可能になる。それも、漢字文献情報処理研究会の仲間たちが開発したNGSM（N-Gram based System for Multiple document comparison and analysis）で処理すれば、複数の文献の処理結果をきわめて見やすい形で一覧表示することができるのである。これは実は大変なことである。古典に関する研究では、コンピュータについては出典探しをする際の検索の便利さが重視されてきたのだが、検索というのは、特定の言葉や表現を、単数ないし複数のテキストの中から探してくれるものでしかない。正規表現を使えば、パターン・マッチングによる柔軟な検索ができるが、基本は同じである。ところが、NGSMの場合は、特定の語句ではなく、複数の文献全体同士を完全に比べ合わせることが可能なのである。そうなれば、その中から、AとBの文献に共通する文字列だけすべて抜き出すとか、AとBとCの文献に共通する文字列をすべて抜き出し、そこからDの文献に含まれる文字列は除く、といった処理が簡単にできるため、従来の検索とはまったく異なる世界が出現するのである。

たとえば、東アジア仏教に決定的な影響を与えていながら、インド選述説と中国偽作説との論争が百年以上続いている真諦訳『大乗起信論』を、思想が近いとされる十部ほどの漢訳経論と比較し、『起信論』の用例が見られる部分を抜き出すと、次のような形で表示される。原文の句読点はすべて外してある。

38　云何以　3　（起：7, 金：31）

このうち、左の38という数字は「云何以」という文字列が、それらの文献に全部で38回出現していることを示し、漢字の右の3という数字は、3字単位で文章を切ったことを表わす。かっこの中は、「起（＝起信論）」に7回、「金（＝菩提留支訳『金剛仙論』）」に31回、用いられていることを示している。そこで、「云何以」の用例が、この二つの文献の中でどのように用いられているかを検索して見ると、「この義は云何？　～を以ての故に（此義云何。以……故。）」といった梵語を直訳した表現ばかりであることが知られる。『起信論』に思想的に近いとされる漢訳経論のうち、中国北地で活躍したインド僧、菩提留支（Bodhiruci）が訳した『金剛仙論』だけがこの表現を用いていること、それもしばしば用いていることが注意されよう。しかも、『金剛仙論』については、訳ではなくて菩提留支の著作であるとする説もあるほか、この菩提留支の系統である地論宗において『起信論』が偽作されたとする説が唐代の頃からあるのだから、この一致は無視できない。

そこで、Bodhiruciの訳であれば「B-」の文字を書名の略号の前に付けるなど、訳者がわかるようにしたうえで、「妄」で始まる2字の

部分の処理結果を表示させると、次のようになる。

16　妄境　2　(起：4, B-十：6, B-金：6)
36　妄見　2　(起：2, B-十：21, B-金：4, G-四：7, R-宝：2)
27　妄執　2　(起：3, B-十：16, B-不：3, P-仏：5)
17　妄取　2　(起：2, B-十：10, B-金：3, R-宝：2)
26　妄心　2　(起：14, B-十：12)
2　妄動　2　(起：2)
7　妄念　2　(起：4, R-宝：3)

「B-十」は菩提留支訳『十巻楞伽経』、「B-不」は同『不増不減経』、「P-仏」は真諦（Paramārtha）訳『仏性論』、「G-四」は求那跋陀羅（Guṇabhadra）訳の『四巻楞伽経』、「R-宝」は勒那摩提（Ratnamati）訳『宝性論』の略である。このような一覧にすれば、『起信論』の語は、真諦の訳経ではなく、菩提留支の訳経の言葉と一致している場合が多いことが一目でわかる。また、14回も用いるほど『起信論』が重視していたらしい「妄心」の語は、菩提留支訳『十巻楞伽経』にだけ12回も見えていることが示すように、とりわけ『十巻楞伽経』との類似が目立つ。『楞伽経』には、ほとんど同じ内容である求那跋陀羅訳『四巻楞伽経』もあって流布していたにもかかわらず、上の一覧では『四巻楞伽経』と一致する文字列は一例しかないことが注目されよう。

このように、NGSMは非常に便利であるうえ、師茂樹氏や近藤泰弘氏が作成して公開している処理ツール（perl用のスクリプトであるため、どのようなパソコン環境でも利用できる）は、きわめて高速である。したがって、学会発表などを聞いていて「おかしい！」と思った場合は、その場でノートパソコンを開いて上記のような処理を試み、質疑の時間になったら、さっと立ち上がって数々の証拠を示しつつ反論する、といったことが簡単にできてしまうのである。ああ、恐ろしい。

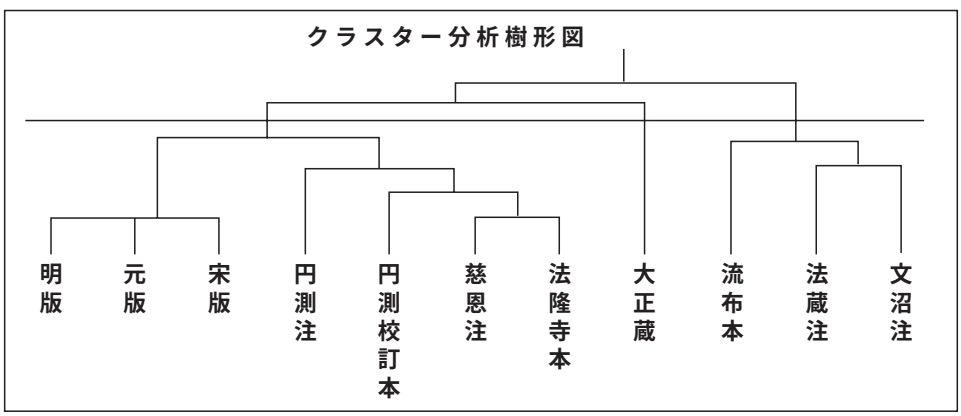
もっとも、研究はここから始まるのであって、NGSMの結果を示すだけでは学問ではない。処理方法に問題はないかどうか確認したうえで、なぜそのような結果が出るのか、多くの資料を活用しながら徹底的に調べ、古代インド人には暗記せよと怒られるだろうが、テキストそのものをじっくり読み直し、考えてゆくのが研究ということになる。いや、そもそも、ある程度の学問的蓄積と柔軟な視点がなければ、NGSMの処理結果を見ても、何も気がつかないだろう。誰かが諸国の遺伝子解析の最新データを盗み出し、筆者にくれたとしても、猫に小判でどうすることもできないのと同様である。電子データを活用するためにこそ、古典研究における基礎的な学力と斬新な発想の力が必要になるのである。

4. クラスター分析による図示

NGSMはきわめて便利だが、図によって示すことができればさらに便利になることは言うまでもない。そこで、師茂樹氏が試みたのが、NGSMの処理結果をクラスター分析してその結果を樹形図として表示させ、諸文献の間の類似度を図示するという実験である。むろん、こうした方法は、村上征勝・伊藤瑞叡氏らの日蓮遺文研究などに見られるように、早くから行われてきた。ただ、それは、コンピュータ、仏教学、文法学などの専門家が協力し、品詞や文の構造などの膨大な情報を付加しつつ、大変な時間と労力をかけて進めるものであった。その点、NGSMによる漢文のクラスター分析の場合は、そうした基礎作業は一切しないため、真偽判定などを行なうための厳密な分析のごく一部の内容を、はなはだ低い信頼度で実行するものにすぎない。ただ、コンピュータにあまり詳しくない文系の研究者でも、漢字文献情報処理研究会の仲間などがフリーで公開しているツールとExcelなどの一般ソフトを用いれば、時間をかけずに簡単にできてしまううえ、雑な処理であるにもかかわらず、なかなか興味深い結果が得られるのが特徴である。

ここでは、有名な玄奘訳『般若心経』の異本の系統を見てみよう。『心経』の異本は無数にあるため、11本だけ比較してみた。クラスター分析の結果である樹形図を見ると、宋版、元版、明版の大蔵経はきわめて似ていて、ひとかたまりになっていることがわかる。次に並ぶのが、円測と慈恩という唐代の法相宗の立役者が書いた注釈中のテキストである。円測が異本として紹介し、こちらの方が正しい訳だとしたテキストは慈恩のテキストにきわめて近いうえ、慈恩のテキストは、法相宗の勢力が強かった奈良の法隆寺が伝統的に用いてきたテキストに近く、いずれも普通のテキストでは「照見五蘊皆空」とあるところを「照見五蘊等皆空」に作っている。つまり、法相宗関係もひとまとまりのグループを形成している。次に来るのが大正新脩大蔵経のテキストである。これは高麗大蔵経を底本としたものだが、これだけが他から孤立した独自なテキストであることは明らかであろう。その右が日本で広く用いられている流布本である。詳しく述べる紙数がないが、これは実はアジアにおいてはかなり特殊な性格を持つテキストである。その右に来ているのは、唐代華厳宗の法蔵の注に見えるテキストである。右端は、敦煌で発見された注釈に見えるテキストであり、著者は不明であったが、福井文雅氏が敦煌写本中の他の断片に「中京招福寺沙門文沼注」と記されているのを発見し、著者を確定された。

この文沼について、福井氏は伝記は不明とされているが、図によれば、法蔵が用いたテキストに近いテキストを用いていることがわかる。そこで、法蔵には文超という弟子がいて著作の逸文が残っており、また敦煌写本は誤写がきわめて多く、字音・字形が似ている場合には特に誤写する場合が多いという点を考えると、文沼は文超の誤写である可能性が出てくる。少なくとも、華厳宗で用いられたテキストに近いテキストを用いていることは確かである。それにしても、日本の流布本はなぜ法蔵など華厳宗が用いたテキストに近いのか。中世に東大寺あたりで出版したテキストが広く流布したのか。



このように、樹形図というのは、見ているだけで実にいろいろなことがわかり、また推測されるものなのである。それだけに先入観念がしみこんでしまう危険性もあるが、逆に、これまで考えたこともなかった事実に気づく場合が多いのが楽しいところである。やがては、多くの研究者が、データを検索するのと同じ程度の気楽さで、上記の比較分析手法を改良したものを参考用に使うようになるだろう。

（駒澤短期大学教授）

なぜぼくのWindowsでミニチュア・アリフが打てないのか？

保坂修司

Windows XPとOffice XP

Windows 2000の登場により日本語システム上でアラビア文字が簡単に利用できるようになってから、ぼくのまわりから徐々にアラビア語版WindowsやOfficeの影が薄くなってきた。昨年XPを導入してからは、基本的なアラビア語の入力・編集はXPだけで事足りるようになったため、今年4月にカイロから帰国するのとほぼ同時にアラビア語のOSは自宅のPCから姿を消してしまった。

ところが4月からひさしぶりに初級アラビア語を教えることになり、少しはまじめに、参考書めいたものでもつくってみるかと思い立ち、東芝Dynabookをまえに、日本語版Windows XP ProfessionalとOffice XPという環境下、アリフ・バー・ターと文字や発音の解説から書きはじめたのである。アルファベット29文字、発音記号と進み、マッダなど特殊な記号のところにきて、ハタと思い出した。そういえば、ミニチュア・アリフってどうやって打つんだっけ？

ミニチュア・アリフって何？

ミニチュア・アリフとはアラビア語アルファベットの最初の文字、アリフを小さくしたもので、これを子音字の上に置くことによって、その字がāという音価をもつことを表す。実際には、その子音字のあとにふつうのアリフをつけるのと同じことなのだが、慣習上、ミニチュア・アリフを使う単語は少なくない。たとえば、「これ」を意味するhādhāのhāがそうだし（هٰذا）、「しかし」を意味するlākinのlāもそうである（لٰكن））。つまりけっこう基本的な単語のなかに用いられているわけだ。その意味では初学者にとって必須の記号といえる。ところが、この基本中の基本がぼくのコンピューターでは打てないのである。いや、打てることは打てる。正確にいえば、ミニチュア・アリフはデフォルトの状態ではキーボードから打つことができないのである。

さて、表1、2は日本語Windows XPとOffice XPのスクリーン・キーボードでアラビア文字の配列を表したものである（上段は日本語の配列）。表2はシフト・キーを押した状態だ。一見してわかると思うが、キーボードのどこを探しても、ミニチュア・アリフはない。淡い期待を抱いてキーを全部押してみても、やっぱりミニチュア・アリフは出てこない。

アラビア文字には子音字しかなく、3つある母音を示すには母音記号が用いられる。しかも、その母音記号なるものは、教育や宗教など特殊な分野を除けば、現代語ではほとんど使われない。もちろん、だからといってミニチュア・アリフがキーボードから抜け落ちてしまった理由にはならない。なぜなら、他の母音記号は初期のころからキーボードで入力できたからである。たしかに、重要度からいえば、短母音記号のほうが大切であろう。ミニチュア・アリフなぞなくとも、まともなアラビア語の知識さえあれば、きちんと単語を読むことは可能だからである。

だが、キーボードから打てないにもかかわらず、このミニチュア･アリフが自動的に画面に出てくることもある。ラーム・ラーム・ハーと叩くと、ふしぎなことに自動的に2番目のラームの上に2重子音を表すシャッダとさらにその上にミニチュア・アリフが現れる。これで「اللّٰه li＋Allāh」、つまり「アッラー（神）のため」の意味になる。このまえにふつうのアリフを置けば、ただの「アッラー اللّٰه」となる。アッラーに敬意を表したのか、正書法にのっとった書きかたに自動変換してくれるのだ。

ミニチュア・アリフ入力のしかた

アッラーと打てば、ミニチュア・アリフが出てくるなら、当然、どこかにそれが隠れていることになる。出しかたはいたって簡単。ちなみにぼくは次のようなやりかたで出すようにしている。

まず［挿入］→［記号と特殊文字］を選択、フォントはふだん使用しているアラビア語フォントを選ぶ。つぎに［ショートカットキー］から適当なショートカットを決める。これだけである（個人的にはArabic Transparentをフォントにし、「Shift＋Alt＋H」をショートカットにしている）。こうすれば、言語をアラビア語に切り替えて、هٰذا と打てば、هٰذا（hādhā）となるわけだ（ただし、あらかじめアラビア語フォントに変更しておいたほうがいい）。

問題なのは、このやりかたがWordでしか使えないことだ。表計算のExcelやデータベースのAccessではこの方法は通じない。どうしてもExcelやAccessでミニチュア・アリフを使いたいなら、いったんWordで入力したものを、セル内に切り貼りするのが手っ取り早いだろう。

ミニチュア・アリフを入力するには別の方法もある。たとえば、現時点でMicrosoftが提供する唯一の統合的UnicodeフォントArial Unicode MSにはいくつかミニチュア・アリフが上に乗っかった文字が収録されている。［挿入］→［記号と特殊文字］でArial Unicode MSを選び、必要な文字を［挿入］すれば、ミニチュア・アリフつきの文字を入力できる。

表1

半 ذ	1 1	2 2	3 3	4 4	5 5	6 6	7 7	8 8	9 9	0 0	- -	= =
q ض	w ص	e ث	r ق	t ف	y غ	u ع	i ه	o خ	p ح	@ ج	[د	
a ش	s س	d ي	f ب	g ل	h ا	j ت	k ن	l م	; ك	: ط	] \	
z ئ	x ء	c ؤ	v ر	b لا	n ى	m ة	, و	. ز	/ ظ			

表2

~ ّ	! !	" @	# #	$ $	% %	& ^	' &	(*	) (	)	= _	~ +
Q َ	W ً	E ُ	R ٌ	T لإ	Y إ	U ‘	I ÷	O ×	P ؛	 >	{ <	
A ِ	S ٍ	D [	F]	G لأ	H أ	J ـ	K ،	L /	+ :	* "	} |	
Z ~	X °	C {	V }	B لآ	N آ	M ’	< ,	> .	? ؟			

（なおこれらの方法は、個人的な経験則にもとづいて導き出したものであり、もっと簡単な方法があるかもしれない。）

なぜ打てないのか

しかし、この小さな字を入力するのになぜこんな面倒な手順を踏まねばならないのか。はじめからキーボードに割りつけておけばいいじゃないか。という疑問が当然出てくる。現在の状況からいえば、少なくともMicrosoftでアラビア文字の処理を担当した人はこの記号が必要ではあるが、キーボードから直接入力するまでの重要性はないと判断したことになるだろう。もう一度、キーボードの配列表を見ていただきたい。アラビア語標準キーボードではもうすでに空いているキーがないのである。それじゃあ、しかたない、といいたいところだが、問題はそう簡単ではない。

第一に、いつからミニチュア・アリフが入力できるようになったのかということだ。きちんと調べたことはないけれども、文字コードの変遷をたどると、見当はつく。実は、現在Windowsのアラビア語標準文字コードである1256にはミニチュア・アリフは入っていない。となると、この字が文字コードに収録されるのは、WindowsがUnicodeに対応してからということになる。したがって、ミニチュア・アリフがデフォルトの機能として入力できるのは、NT系は2000以降、9x系はXPになってからといえるだろう。

アラビア語文字コードの流れ

アラビア語文字コードはかならずしも順調な進化を遂げたわけではない。1980年代には20以上のコードが乱立していたという。アラビア語を母語とする国が多すぎるのがそもそもの原因だが、それらの国が統一コード制定に向け協力関係を築けなかったことも無視できない。しかも、アラブ諸国間でさえ協力できないのだから、アラビア文字を使う（あるいは使っていた）国ぐにとの連携などSFの世界である。そのため、アラビア語文字コードはつねに迷走しており、それは現在までつづいている。

最初のアラビア語標準文字コードは1981年のCUDAR-Uだといわれている。その翌年、ASMO-449が制定された。これは1986年にASMO-708となり、8bitに進化する。これが国際標準となり、ISO-8859-6となる。ちなみにMacintoshのアラビア語はこのISO-8859-6にもとづいている。

現在アラブ諸国でもっとも幅を利かせている文字コードはこの国際標準といいたいところだが、現実にはWindowsの1256がデファクト・スタンダードとなっている。一私企業（しかもアラブ諸国ではなく、アメリカ企業）のつくった文字コードがアラビア語PCの標準になっているわけだ。これはアラビア語にとってはきわめて不幸であり、かつ危険なことでもある。しかも、アラビア語はそれでもなお複数のコードが統合されないまま並存しているのだ。XPとなって、Unicodeが統合への方向性を示してくれたが、レガシーの問題を含め、包括的解決までの道のりは長く険しい。またはたしてUnicode制定にあたって、どれだけアラブ諸国のイニシアティブが発揮できているかも疑問である。同じことはアラブ諸国の標準化機構であるASMO（Arab Standards and Meteorology Organization）でさえ指摘できる。

ASMO制定の文字コードがなぜアラブ諸国で一般化しなかったのか。最大の理由はアラブ諸国のコンピュータ事情、つまり、アラブ世界のコンピュータがソフト・ハード両面でつねに外国主導だったことにある。PCの世界でいえば、DOS時代はIBMが独自の文字コードを使用、DOS後期からWindows時代にはMicrosoftの規格が圧倒的優位を保つ。この時期、クウェートを本拠とするSakhr社（現エジプト）がDOSおよびWindowsのアラビア語化ソフトで名を馳せたが、結局Microsoft側からの激しい攻勢を受け、この分野から撤退せざるをえなくなった。Sakhrでアラビア語化を担った連中の多くはMicrosoftに引き抜かれたり、湾岸危機で会社を離れるなどしたため、アラブ系企業による文字コード関連イニシアティブは急速に減退することになった。一方、Appleは国際標準を採用したが、いかんせん、Macintoshのアラブ世界での普及率の低さにより、国際標準の拡大にはほとんど貢献できなかった。

文字コードとミニチュア・アリフの問題は、もう1つ重要な問題を含む。ミニチュア・アリフがきちんとコード化されたのがUnicodeからであったなら、なぜそれまでこの文字が採用されなかったのかという問題である。現時点でこそキーボードは埋まっているが、少なくとも98時代まではアラビア語キーボードは、シフトを押した状態では、空白だらけだったのである。Unicodeが採用されて、キーボードのあまっている場所を埋めていくときに、なぜそれほど重要とも思えない記号や母音記号をつけるだけで処理できそうなリガチャーが採用され、ミニチュア・アリフが採用されなかったのか。はたして、各アラブ諸国の宗教関係者や教育関係者はそれでOKを出したのか。文字コード制定やキーボードの設定にどれだけ彼らが関与できたのか。キーボードが埋まっているといったが、拡張キーボードを採用することだってできたはずだ。即断できないが、この問題には、アラブ世界のコンピュータ事情、とりわけハードウェア・メーカーの決定的欠如とソフトウェア開発における文科系的要素の不在が根深く関係しているような気がしてならない。

おわりに── Unicodeの問題点

Unicode導入でミニチュア・アリフ問題は解決されたようにみえる。だが、そうは問屋が卸さない。Unicodeのアラビア語にはまだ問題が残されている。たしかにUnicodeによって、アラビア語のみならずアラビア文字を使う言語の一元的な処理が可能になった。しかし、細部を見渡してみると、きわめて問題が多いことがわかる。たとえば、ミニチュア・アリフがらみでいうなら、この字を含むアラビア語母音記号の扱いがUnicodeではかならずしも明確ではない。これらの記号は、Unicode Standardではいわゆるnonspacing marksと分類される。しかし、Unicodeにおいては、記号なしの場合、1つの記号がついた場合、1つ以上の記号がついた場合で、検索、並べ替えの基準が明示されていない。MicrosoftのOffice系ソフトではかろうじてWordのみが母音記号の有無を区別できるが、他のソフトでは検索・並べ替えとも不正確である。

もう1つの大きな問題はリガチャーの扱いである。キーボードからはラーム＋アリフがらみのリガチャーを入力することができる。またUnicodeでは、それ以外多数のリガチャーが用意されたが、大半はその意図がわからないものばかりで、実用的とはいいがたく、しかも真に必要なリガチャーがほとんど採用されていない。しかも問題なのはリガチャーに個別のコードが割り振られ、同じ文字の組み合わせでありながら、異なる文字として認識されてしまうことである。たとえば、hādhā の場合、ﻫﺬﺍ（ミニチュア・アリフは前述のように［Shift＋Alt＋H］で入力している）と［記号と特殊文字］から挿入したArial Unicode

MSのﻫﺬﺍはWindows上では異なる単語として認識される（後者のhā ﺍ がリガチャー）。

とまあ、WindowsやUnicodeの問題点をミニチュア・アリフを例にして挙げつらねてきたが、それでもUnicode以前と比較すれば、着実に進化している。Unicodeに含まれたコーランを表記するための記号類を見れば、完全とまではいえないにしろ、イスラームの専門家の意見が大幅に採り入れられたことがわかる。今後はUnicodeがよりいっそう洗練され、さらにそれが正確にソフトウェアに反映されていくことを期待したい。

（防衛大学校講師）

クメール文字コードをめぐる諸問題

原田至郎

メールやウェブ・ページを開いたときに意味不明の記号が羅列される「文字化け」という現象をご覧になったことがある方は多いと思います。コンピュータは、データの効率的な処理のため、文字を数値（コード値）に置き換えて扱いますが、データに用いられている文字とコード値の対応ルール（文字コード）が、それを見るためのソフトウェアが想定しているものと一致しない場合、「文字化け」が起きます。実は日本では、大別して3種類の文字コードがよく用いられています。幸い、コード値を与えられる文字の範囲（文字集合）は基本的には一致しており、コード値の与え方（エンコード）が異なっているだけなので、多くのソフトウェアでは自動判定・自動変換を行っているのですが、ときどき失敗すると「文字化け」が現れるのです。

カンボジアの国語はクメール語で、その表記にはクメール文字を用います。これは、子音文字と母音記号を組み合わせて表記上の音節を構成する結合音節文字です。古代インドのブラーフミー文字を祖としており、子音文字は内在母音を含んでいて単独でも音節を構成できるという、インド系の特徴も受け継いでいますが、既に1400年以上の独自の歴史があり、声門閉鎖を表す子音文字や、内在母音の異なる2系統の子音文字の存在など、ユニークな特徴も備えています。

カンボジアでも、クメール文字で書かれた新聞や雑誌、様々な文書の作成にコンピュータが利用されています。しかし、いざデータ交換をしようとすると、そこには常に「文字化け」の問題がつきまとっているのです。カンボジアでは文字集合もエンコードも異なる文字コードが乱立しているからです。

カンボジアで使われているOSは、圧倒的にWindowsです。しかし、タイ語版Windowsのある隣国とは異なり、ローカライズされたものはなく、英語版が使われています。2000やXPでは多言語サポートが進んでいますが、クメール語はまだ含まれていません。

英語版OS上でどうやってクメール文字を使っているのでしょうか？簡単に言えばOSをだますのです。コード値に関連付けられた文字の図形表現の集合を収めているのがフォント・ファイルです。これを改造するツールを使って、英語版OSに標準装備されているラテン文字フォント・ファイルの中のラテン文字の図形表現をクメール文字の図形表現と置き換えて作った新たなフォントを使えば、OSはラテン文字を表示しているつもりでも実際にはクメール文字が表示されていることになります（図1）。綺麗なプリントアウトを得るだけなら、これで一応用は足ります。

しかし、この方法には様々な問題があります。まず、クメール文字とラテン文字がコード値を共有していますので、テキスト･ファイル化などにより、ひとたび使用フォントの情報が失われてしまうと、両者の区別がつかなくなります。また、必要な図形表現の数が多いため、英語圏ではあまり使われない文字の領域にまで割り当てなければなりませんが、そのような文字は、例えば1文字打つのに4つのキーを押さねばならないなど、かなり煩雑なキーストロークを要することもあります。実際にはそれでも領域数が足りず、正しくない字体で我慢せざるを得ない場合もあり、そのような印刷物の氾濫は伝統的な文字文化へも影響を与えつつあります（図2）。

ここで、抽象概念としての「文字」とその「図形表現」（グリフ）が区別されることにご注意ください。同じ文字でも、状況によって表示位置や長さが変わることがよくあります。隣の文字と結合して合字を作ることもあります。また同じグリフでも違う文字を表している場合もあります。さらに、文字の順序とグリフの順序が逆になる場合すらあります。検索や並べ替えなどデータ処理のしやすさから言えば、グリフではなく文字を処理単位とするのが望ましいのですが、そうなると文字の列から適切なグリフの列を作る処理が表示段階で必要になります。既存の英語環境にはそのような仕組みはないため、これらのフォントでは、文字ではなくグリフをラテン文字に割り当てているのです。その結果、本来同じ文字であっても、位置や長さ、形が違うものはすべて別のコード値を持ち、別のキーストロークで入力することになります。

一番深刻な問題は、クメール文字のグリフとして何を取り上げるのか、またどのラテン文字にどのクメール文字を割り当てるのかというルールが統一されていないことです。その結果、ある系列のフォントを使って作成した文書を別の系列のフォントを使って読もうとすると「文字化け」してしまうことになります（図1の最右列）。フォントの系列は多数あり、比較的有力なものだけでもLimon、abc、KHEKなどが使われていて、事実上の標準の確立にも至っていません。長く続いた国内の政情不安の影響もあり、国家標準もありません。このことは、様々な媒体を通じたデータ交換を困難にし、ITのメリットを享受する上でも障害となっています。

Khmer Philology Project（KPP）は、このような問題に取り組むことを目的として、石澤良昭・上智大学教授をリーダーとし、財団法人アジア太平洋研究会を事務局として、1999年3月に発足した非営利プロジェクトです。複数のカンボジア出身者も参加しています。本来畑違いの私もたまたまこれに参加することになりました。現地調査を行い、現状の問題点を解決する方針を立て、カンボジア出身メンバーが中心となって、理想的なクメール文字コード表を作成しました。これは、正書および実用の観点から必要なクメール文字とグリフを網羅し、両者を明確に区別しながら、コード化したものです。そこで私は、その実用性を検証するツールとして、Intelligent Khmer Writing System（IKWS）を開発しました。これは、体系的で容易な方法で正しいクメール文字を打つためのWindows用入力支援ツールです。2000年1月にプノンペンで開かれた国際クメール学会での成果報告をきっかけに、KPPはカンボジア政府、王立アカデミ

図1　ラテン文字フォントを改造したクメール文字フォントの例（「クメール」の入力例）

フォント名（系列）	コード値（10進数）	ラテン文字	クメール文字	フォント交換時
Limon S1 (Limon)	69 120 181 114	Exµr	ខ្ញុរ	ខែឡូរ
KhmTimes ~~(abc)~~ lkl	69 120 218 114	ExÚr	ខ្ញុរ	ខែ្ឍរ

図2　正しくない字形の例

គាត់ （彼、彼女）　正しくは、គាត់ （最初の字の上部の山の数に注意）

ー、現地NGOなどと、国家標準文字コードの制定に向けて協力していくことになりました。様々な意見交換や、IKWSを使った実地検証を経て、標準文字コード草案が固まっていきました。

ところが、実はカンボジア政府も知らないところで、クメール文字コードの国際標準が決定していました。2000年9月に出版された、世界中の文字を統一的にコード化する国際標準ISO/IEC10646-1（Universal Multi-Octet Coded Character Set略してUCS）の第二版に、クメール文字ブロックが新たに追加されていたのです。追加すること自体は、この国際標準を担当するISO/IEC合同技術委員会第2小委員会第2作業部会（ISO/IEC JTC1/SC2/WG2）での審議を経て、既に1999年9月締め切りの投票で最終決定されていました。

世界中の文字のコード化と聞いて、Unicode（ユニコード）を思い浮かべる方もいらっしゃるかも知れません。これは正式にはThe Unicode Standardといい、The Unicode Consortiumという業界団体が制定しているものです。最近のWindowsやMacOSなど主要OSの内部処理にはUnicodeが用いられており、その意味で業界標準と言えますが、Unicode自体は国際標準ではありません。しかし、取り決めによって、UCSとUnicodeは整合性を図ることになっており、文字の名前とコード値の対応関係は両者で一致しています。JTC1での最終投票の直後に出されたThe Unicode Standard Version 3.0には、国際標準の出版より一足早く、クメール文字ブロックが追加されていました。

しかし、UCS/Unicodeのクメール文字コードには、数々の問題がありました。その最たるものは、カンボジア人が口を揃えて指摘する「脚がない！」ということです。

クメール文字では、既に述べた通り、子音文字に内在母音が含まれています。これが問題となるのが、子音結合(consonant cluster)を表す場合です。例えば、クメール語で女性はstreiと言いますが、単にs、t、rに対応する子音文字とeiを表す母音記号を並べただけでは内在母音が残った複数の音節になってしまうのです。この場合、クメール文字では、最初の子音s以外のtとrを通常の子音文字とは別の子音記号で表すことで、全体が子音結合であることを表します。この子音記号は、基本的に子音文字の下に付くので、「脚(coeng)」と呼ばれるのです（図3）。

ところで、この子音結合の表記は、内在母音を持つインド系文字に共通の問題ですが、解決方法は一通りではありません。例えば、デーヴァナーガリー文字では、内在母音の不在を子音文字の一部字画の省略で表したり、子音結合全体を特別の字体で表したり、あるいは子音文字に内在母音を抑制する記号を付加した上で並べたりします。同じ子音結合を表記するのに、複数の方法が存在するのです。Unicodeのデーヴァナーガリー・ブロックでは、「文字」列データ中では常に子音文字と内在母音抑制記号(ハル記号、UCS/Unicodeではvirama)を用い、字画省略字体や特別字体は独立の「文字」とは見なさず、必要な際には表示段階で上記文字列をこれらの「グリフ」列に置き換える方法（virama model）を採用していました。さらに、インドで使われる文字については、タミル文字やカンナダ文字などかなり構造の異なるものまで含めて、この方法が適用されました。これは、Unicodeが参考にしたインドの国家標準の方針を引き継いでいます。

クメール文字がUCS/Unicodeに追加されたときにも、同じインド系だからという理由で、この方法が前提とされていました。すなわち、「脚」は子音文字の「グリフ」のひとつに過ぎないと見なして文字コード値を与えず、データ中で内在母音抑制記号と子音文字が連続したら表示プログラムにこれを「脚」の形として表示させればいい、としたのです。

しかし、クメール文字にとって、「脚」は子音文字とは別個の子音記号です。規則上は、これを順に下に配置していくだけでどんな子音結合でも表せ、変形を伴うデーヴァナーガリー文字とは構造に差異があります。しかもこれが唯一の方法であるため、viramaのような汎用の内在母音抑制記号はクメール文字には存在せず、従って本来virama modelの採用は不必要かつ不可能なのです。さらに、「脚」は音節末子音を表すために使われることもありますが、この場合は直前に母音があるわけで、内在母音抑制記号を置くことは機能面からの理屈にも合いません。しかし、UCS/Unicodeは、実際には存在しない架空のvirama相当文字(COENGと命名)の創作までして、virama modelを採用したのです。

このことは、母音記号の認定方針など他の問題と相まって、データ長と必要な処理の増加という非効率性をもたらしています（図4）。なぜこんなことになったのでしょうか？議事録からは、実質審議におけるクメール文字のネイティヴや専門家の不在、そして上記の諸点に関する無関心が読み取れます。提案者とカンボジア人研究者らとの個人的な接触が提案前にはあったのですが、子音文字とは別に「脚」が必要であるという彼らの主張は、審議の過程で退けられました。カンボジアにはISOの購読メンバーとして登録されていた部局が

図3　子音結合と「脚」

ស្ត្រី [strei]（ស sa+◌្ coeng ta+ ◌្ coeng ro+◌ី ii）cf.ស៊ាំ [sɔ:tri:](sa+ta+ro+ii)

図4　カンボジア案とUCS/Unicode案の対比の具体例（「私」）

	カンボジア案	UCS/Unicode
ខ្ញុំ [khɲom]	ខ+◌្+ញ+◌ុ+◌ំ（3文字） kha+coeng nyo+om	ខ+COENG+ញ+◌ុ+◌ំ（5文字） kha+COENG+nyo+u+NIKAHIT

あったのですが、審議の過程で何の照会もなく、WG2の議事録からは公式の照会を意識的に避けたことすらうかがえます。

国内標準を国際標準に整合させることを求めるWTO-TBT協定は、同時に、国際標準の決定に際し、全関係者の、特に途上国からの、参加を確保する努力を求めています。一定の強制力を持つ国際標準をより良いものにするためには当然必要なことですが、残念ながらこの精神は尊重されていませんでした。

いずれにせよ、できてしまった国際標準を放置したまま国家標準を決めることは、新たな非互換性を招きかねません。結局、カンボジア政府は国際標準の改訂を求めて、2001年5月にJTC1、ISO、IECなどに抗議状を送り、10月のWG2会議、2002年2月のUnicode Technical Committee (UTC)、5月のWG2会議に代表団を送り込むとともに、この間に対案を含む多くの上申書を提出しました。KPPもカンボジア政府担当者から依頼を受け、これにできる限り協力しました。

この結果、少なからぬ関係者が現行標準の問題点に関するカンボジア側の主張に賛同してくれました。しかし、残念ながら、WG2の主流派は、決定過程に関する議論を許さず、現行標準が非効率ではあっても何とか使える以上、その安定性を維持せねばならないとして、「脚」や欠損母音記号の追加を拒否しました。追加しても安定性は損なわれないという意見もあったのですが、一部関係者による舞台裏での権謀術数などもあり、理詰めでは済まない世界であることが痛感されました。

一方、WG2と一部メンバーが重なるThe Unicode Consortiumは、現行標準の安定性には固執しつつも、その内容および決定過程における過ちについては明確に認め、「詫び状」をカンボジアに送りました。結局カンボジアは、2002年5月、WG2が過去の過ちを認め、同様の悲劇の再発防止策を講じることを条件に、「脚」や母音記号の追加要求は諦め、独自の国家標準も作らず、UCS/Unicodeを受け入れる、という苦渋の決断を下したのです。

残念ながら上記の条件は、WG2主流派によって、未だ公式には無視されています。しかし、カンボジア側の取り組みが無駄であったわけでもありません。実際の審議がより慎重に行われるようになったことは感じられますし、標準の内容についても、先の二つの大きな問題以外の数々の不足や誤りについては、カンボジア側からの提案に基づいて修正手続きが進んでいます。さらに、次に出版されるThe Unicode Standard Version4.0のために、クメール文字コードの正しい理解と実装のための解説文をカンボジア側から寄稿しました。これらは、一連の動きがクメール文字問題への関係者の関心を高めたことと相まって、正しいクメール文字サポートのより早い充実に大いに寄与するものと期待されます。

カンボジアの人々が、真の意味で、自国語・文字でITを活用できる日が一日も早く来ることを願っています。

（大学院情報学環助教授）

センター便り

平成14年度漢籍整理長期研修

昨年度に引き続き、本年度も漢籍整理長期研修が実施された。研修期間は夏休みをはさんで前期と後期の2期に分けられ、前期は6月24日から7月5日まで、後期は9月30日から10月4日までであった。参加者は、大学図書館の司書8名、院生4名の合計12名であった。講師としては、東洋文化研究所のスタッフのほかに、所外から、京都大学の井波陵一教授、宇佐美文理助教授、富山大学の藤本幸夫教授、日本女子大学の陳捷講師、鹿児島大学の高津孝教授、慶応義塾大学の高橋智助教授、宮内庁書陵部の横山謙次氏と安藤清氏、国立公文書館の長澤孝三研究官、東洋文庫の中善寺慎司書、の諸先生方にご助力を賜った。また東洋文庫と内閣文庫には見学の便宜を図っていただいた。ここに記して謝意を表させていただく次第である。

5センターセミナー

今年の「全国文献・情報センター人文社会科学学術情報セミナー」（略称5センターセミナー）は、11月18日(月)と19日(火)神戸大学大学院国際協力科で、また同22日(金)に東京大学山上会館で開催される予定である。

同セミナーのプログラムについては、本年度主催センターである神戸大学経済経営研究所附属政策研究リエゾンセンターのホームページ（http://www.rieb.kobe-u.ac.jp/doccenter/guide.html）を参照されたい。

明日の東洋学
東京大学東洋文化研究所附属東洋学研究情報センター報 第8号
発　行　日　2002年10月31日
編集・発行　東京大学東洋文化研究所
附属東洋学研究情報センター
〒113-0033　東京都文京区本郷7丁目3番地1号
電話　03-5841-5839（直通）
FAX　03-5841-5898
ホームページ
http://www.info.ioc.u-tokyo.ac.jp/

デザイン　コスギ・ヤエ／印刷　㈱ヒライ