

電子文書における多言語組版の標準化 —事例調査による要求抽出—

上村 圭介 *

要旨

国際的な文字コードの開発により、ある言語で使われている文字の一つ一つをコンピューターの内部処理で使用する上での問題は少しずつ解決しつつある。次に目指すべき課題は、言語固有の組版やレイアウトの規則に基づいて、一つ一つの文字を適切に配置するために必要なスタイル指定の開発である。

英語、ヨーロッパの言語、日本語の組版のためのスタイル指定の記述枠組みは、国際的な規格としても制定されている。しかし、その規格が、その他の、特にアジアの諸言語に対して適用できるかどうかはこれから検証する必要がある。その結果次第では、アジアの諸言語で用いられている組版要素や特徴の規格への追加が必要となる。

本稿では、このような視点から行なわれたアジアのいくつかの言語における組版の現状調査の結果について報告し、多言語組版の標準化への示唆について考察する。

Document Style Requirements from a Perspective of Multilingual Document Formatting

Keisuke Kamimura

Abstract

An increasing number of languages can have their writing in coded characters as standardisation efforts were carried out at international levels. However, in order for any electronic document to be accepted fully by the user, a speaker of the language, there are other issues to be answered than the coded character.

One of these issues is document style specification. Although there has been international initiative towards the standardisation of document style specification, the applicability of the standards must be examined in light of languages other than those that were originally presumed in the standard, particularly to that of languages in Asia.

This article introduces the reader to the results of the survey in document style and layout in some Asian languages, and draws implication to the future standardisation of multilingual document formatting.

1 はじめに

電子的コミュニケーションの手段が社会活動に浸透するにつれ、電子文書による情報交換の

重要性は増している。中でもマークアップされた電子文書の重要性は、文書の再利用性の高さなどから、とりわけ高まると考えられる。マークアップ文書は、通常処理の過程で人間可読(human readable)な文書へのフォーマットが必要となり、その際、文書構造に基づき、元の文書から目標文書への変換方法を示したスタイ

*学校法人国際大学グローバル・コミュニケーション・センター (GLOCOM) / Center for Global Communications, International University of Japan

ルとレイアウトの記述（スタイル指定）が必要となる。

電子文書のスタイル指定の国際的な標準としては、ISO/IEC 10179 “Document Style Semantics and Specification Language”、W3C Recommendation “Extensible Stylesheet Language (XSL) Version 1.0” などが制定されている。しかし、これらは電子文書のスタイルとレイアウトを記述するための枠組みを規定するものであって、実際に電子文書のフォーマットを行なうには、その枠組みに基づいて記述したスタイル指定が必要となる。

日本語組版、欧文組版に必要な文書スタイルをまとめたライブラリとしては、すでにISO/IEC TR 19758 “DSSSL library for complex compositions” があるが、それ以外の言語で記述される電子文書が適切な文書フォーマットによって整形・組版されるためには、その言語に求められる組版要素や特徴を抽出し、スタイル指定言語に置き換えなければならない。ここに新たな技術開発や標準化の必要性が生じている。

2 問題の所在

2.1 電子文書の新たな課題

コンピューターやインターネットでの言語利用の問題を考える上でもっとも重要なのは、文字コードの問題だろう。コンピューターやインターネットで、ある言語を利用するためには、何よりもまず、その言語で利用される文字や記号類を表現するのに十分な符号化文字が与えられなければならない。

南アジア、東南アジア、あるいは中央アジアの言語の表記に使われている文字では、文字の同定に際してそれぞれ異なる困難があり、電子的にその言語の表記に使われている文字を表現する段階で解決すべき問題を抱えている言語も

ある。また、その文字コードを操作するための基本および応用ソフトウェアも当然に必要となる。そのため、多言語処理という観点からは、文字の符号化について関心が集まるのがこれまで多かった。

しかし、電子文書の利用の上で必要な標準化は文字の符号化だけではない。むしろ、その言語によるコミュニケーション活動が電子文書に基づいて進められるにつれ、さらに上位の問題が顕在化してくると思われる。本稿では、多言語組版、あるいは文書フォーマットは、そのような問題の一つであり、今後の標準化の大きな課題となるだろう。

なお、本稿の議論では、便宜上「組版」、「レイアウト」、「スタイル」、「フォーマット」という言葉を次の意味に使用する。

組版 流し込むものの組み立て。多くは文字組み。

レイアウト 流し込み先の領域の記述、または関連した領域における配置処理。

スタイル 個々の要素に目的とする組版を行うために記述する性質、または特質の集合。

フォーマット フォーマットティング。組版に関連する処理。

また、本稿では、文書フォーマットのもっとも基本となる単位を総称して「文字」と読んでおく。そのため、その意味は必ずしも文字符号化の議論での文字の定義と合致しない。

2.2 文書フォーマットと多言語処理

文字、あるいはフォントは、本来ある程度、流し込むものの組み立てについての情報をもつ。そのため、いわゆる「べた組み」によっても、ある程度文書フォーマット上の目的を達成することができる。可変幅の書体を利用することで文字同士の間隔が調整できるということは、べた組みによっても、文字がもつ属性情報を利用

することで文書フォーマット上の目標の一部が達成可能であることを示している。

しかし、べた組みにおける文字の処理は、ある文字が置かれた環境（つまり、前後に配置されている文字の状況）にもとづいて行なわれる。ところが、ある文字が行末に出現できるかどうかは、文字的な要素だけでは決定できない。ある位置に句読点（punctuation）が出現できるかどうかは、その句読点の特質だけを見ることである程度決定できる（厳密には日本語の場合には、句読点の前後に出現する語の語彙的あるいは形態的な特徴を考慮しなければならないが）。一方で、単語間の区切りや単語内の区切りは、文字だけを見ても決定できず、一般的には**語彙的に**判断しなければならない。ハイフネーションに辞書が必要というのは、文書フォーマットに語彙的判断が必要ということを意味する。

一方で、一つの行の内で処理が完結しない文書フォーマット上の特徴もある。例えば、段落やページといったより大きな単位に対して行なわれるフォーマット処理が挙げられる。また、段落の字下げや行揃え、段落幅や段落間スペースといった特徴は、ページの中の位置関係に関連して決定される。一つの行の内部で処理が決定できないというのはそういう意味である。このように、文書フォーマットには、文字から語、語から行、行から段落、段落からページといういくつかの階層的な構造を考えることができるが、それぞれのレベルにおいて行なわれている規則や慣習の総体が文書スタイルである。このようなレベル分けの考え方は、最近の組版・文書処理システムであるΩ（オメガ）などには取り入れられているようである [1]。

これらの文書フォーマットに関する特徴は、言語ごと、あるいは文化ごとに異なる。もちろん、書きことばという制約、文字という記号を

版面に配置するという制約から、言語ごとに共通する特徴も少なくないが、言語固有の特徴もあるはずである。これらの特徴を電子文書に反映させるためには、実際の書きことばの慣習に基づき、どのような特徴があるのかを明らかにしなければならない。

2.3 電子文書の三類型

本稿で対象とする電子文書の形式は、書きことばを表現するためのものであり、最終的にはどれも人間が閲読することを前提にしている。機械可読文書は電子文書としての構造をもつが、第一義的には書きことばとして人間の閲読に供されるものではないため、対象とはならない。

人間が読むための形式ということに限定すると、現在使われている電子文書には、次の三つの類型を考えることができる [2]。

最終出力形式 最終的な出力のために用いられる形式。これ以上編集を施さないことが前提の文書形式である。書かれた内容についての構造を保持するとは限らない。

PostScript、DVI、PDF などが相当する。

編集可能形式 最終的な出力に使われることもあるが、編集を前提とした文書形式。編集操作は、アプリケーション・プログラムと結び付けられており、多くの場合、このタイプの文書形式は、ソフトウェア・ベンダーの独自仕様となる。Word（マイクロソフト）、一太郎（ジャストシステム）などの形式がこの類型に相当する。

構造記述形式 構造記述をされた文書と、その構造に対応したスタイル情報を別々に与えることができる文書形式。SGML と DSSSL の組み合わせや、XML と XSL の組み合わせがこれに相当する。

本稿が対象とするのは、このうちの「構造記

述形式」に属する電子文書である。この類型に属する電子文書は、文書構造に基づく論理的なマークアップと、論理的マークアップに対応したスタイル情報を必要とする。「最終出力形式」および「編集可能形式」の類型では、文書構造が明示化されないまま、最終的な見た目をもっとも重視して文書のフォーマットが行なわれる。一方で、「構造記述形式」の文書では文書構造とそれに基づくスタイル情報が明示的に与えられなければならない。

2.4 デジタル・デバイドからみた多言語組版

ところで、このように電子文書の利用における新たな課題であるスタイルとレイアウトの問題は、情報化がもたらす格差という点から見た場合、どのように位置付けられるのだろうか。スタイル指定は、電子文書における文書フォーマットを考える以上、どのような言語であっても必要となるものだが、その言語の話者数や、その言語集団がもつ経済力などによって整備状況は変わってくる。これも、「デジタル・デバイド」問題の新たな側面であると言える。

国際的な文字コードの開発などにより、コンピューターの内部表現レベルでの文字処理の問題は少しずつ改善している [3]。しかし、電子文書の普及という流れの中で見た場合、符号化を与えられた文字だけが電子文書に必要な道具ではない。次のステップとして、言語固有の組版やレイアウトの規則に基づいて、一つ一つの文字を適切に配置するために必要なスタイル指定の開発を検討する必要がある。

電子文書のスタイル指定という問題は、情報化が進んでいる国においても、これまで一部の専門家を除けば強い関心の対象ではなかったと言え、途上国では殊更その整備は国際的な支援が必要な分野であると言えるだろう。

また、このような形の支援を行なう上では、受け入れ国側の事情にも配慮する必要がある。

スタイル指定という分野は、純技術的に解決が困難な課題というよりは、言語・文化的な視点から、どのようにその言語の書きことばを電子文書の世界に反映させていくのかという課題である。そのため、より直接的に言語話者の言語行動に関わってくる。そのため、スタイル指定には、実際その言語で行なわれている組版の状況を反映させる必要がある。

3 文書スタイルの調査

3.1 調査の位置付け

本稿は、財団法人国際情報化協力センター (CICC) が実施した「多言語文書情報交換に関する国際規格共同開発調査」の一環として設置された「アジア文書情報交換委員会 (Asian Document Style Information Interchange; 以下 DocSII 委員会)」によって実施された調査研究の成果に基づいている。「多言語文書情報交換に関する国際規格共同開発調査」は、財団法人日本規格協会が経済産業省から委託を受け実施した「国際規格共同開発調査事業」のテーマの一つであり、CICC が日本規格協会からの再委託により実施したものである。筆者は、2002 年より DocSII 委員会に委員として参加し、アジア諸国の組版や電子文書についての現地調査やヒアリングを行なった。

3.2 調査の目的

文書スタイルは言語ごと、文化ごとに異なるものであると考えられるため、その言語の書きことばの慣習にあわせた文書スタイルの要素や特徴を抽出する必要がある。

DocSII 委員会では、アジア諸国の代表的な言語の書きことばがもつ文書スタイル上の特徴を、その言語の書きことばの実態、とりわけ雑誌、書籍などの出版物の状況をもとに選びだし、その言語による組版の専門家などとの議論を踏まえ、文書スタイル規格への要求として抽

出することを目的としている。抽出された文書スタイル規格への要求は、既存の文書スタイルの関連規格と対照し、必要な追加提案をするための根拠として活用される。

DocSII 委員会は、日本規格協会情報技術標準化センター（INSTAC）の電子出版技術調査委員会 WG2、ISO/IEC JTC1/SC34/WG2 などと、各言語からの要求を国際・国内規格に反映させるべく協力を進めている。

なお、DocSII 委員会の活動は、前述の通り、国際的なデジタル・デバイド解消というもう一つの目標をもっている。電子文書の処理を言語に依存せず透過的に行なえることが、国際的な（あるいは言語間での）デジタル・デバイドを解消するために必要である。本稿のもとになっている調査研究は、このような視点からアジアの諸言語での電子組版を可能にするための基礎的な調査研究を行ない、デジタル・デバイドの解消に寄与しようとしたものである。

3.3 調査の方針

DocSII 委員会では、アジアの言語を次の三つのグループに分けて、調査活動を進めている。

- 漢字系文字圏（日本語、韓国・朝鮮語、中国語で使われる文字）
- タイ系文字圏（タイ、クメール語で使われる文字）
- インド系文字圏

今年度中に、漢字系文字圏とタイ系文字圏については、スタイルとレイアウトに関する固有の要素や特徴の取りまとめを完了できる見通しである。また、インド系文字圏については、今年度から基礎的な資料の収集を始め、来年度以降改めて固有の要素や特徴についての抽出を計画している。

4 各国の組版の検証

DocSII 委員会が、これまでに文書スタイルの調査の対象としてきたのは、中国語（簡体字）、韓国・朝鮮語（ハングル）、タイ語（タイ文字）、ラオス語（ラオス文字）、タミル語（タミル文字）である。そのうち現地調査を行なったのは、タイ語、ラオス語、タミル語の三つであり、ほかの言語については収集した参考資料の分析と、訪日した電子文書関係の専門家との議論により個別の文書スタイル上の特徴について検討した。

4.1 いくつかの例

以下に、筆者および DocSII 委員会の調査により精査を要すると認められた事例の中から、タイ文字における事例を紹介する。これらの事例は、和文あるいは欧文組版の観点から‘foreign’であると思われた要素だが、当然現地では慣習的なものであると考えられる。なお、図表の名称には収集した資料の整理番号をそのまま使用している。

非母語話者の視点からは、‘foreign’あるいは‘exotic’に見える要素にばかり注意が向いてしまうことがあるが、実際にその言語の電子文書を使用する上では、非母語話者が考えるほど重要な要素ではないことがある。多言語の事例を収集・分析する上では、視点の片寄りが起こらないよう十分配慮するだけでなく、その言語による組版の専門家との意見交換をすることが不可欠である。

4.2 スタイル上の特徴

スタイル上の特徴として興味深いのは、ボールドが二重に適用されたケースだろう。図 1 は、ボールドのテキストの中に、さらにボールドが適用されたテキストが入れ込まれている例である。

もう一つは、下線の振る舞いである。図 2 に

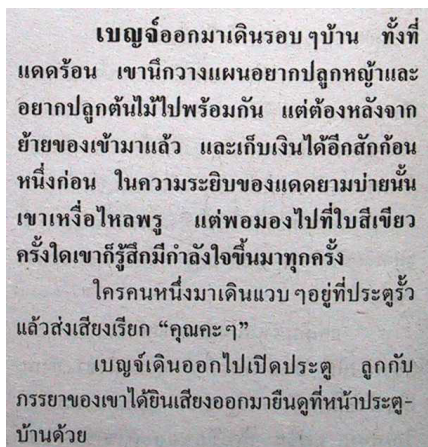


図 1: 例 0103

示すように、下線は、基準線よりも下に突き出た文字を避け、次の文字の位置から再開する。ちなみに、現地でのヒアリングによると、タイ語では学校教育で文字を習う際にもこのような下線の用法を学ぶとのことであった。

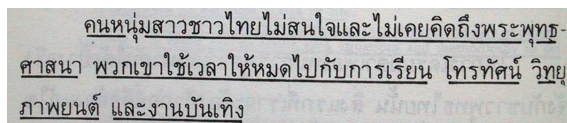


図 2: 例 1002

4.3 レイアウト上の特徴

図 3 は、見出しの開始位置の典型的な例である。インデントされた本文の第一行と同じ水平位置から見出しが開始している。図 3 は章番号付きの見出しであるが、章番号がない見出しであっても、本文の第一行と同じ水平位置で開始する例が多く見られた。

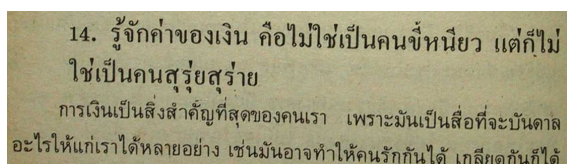


図 3: 例 0903

タイ語、ラオス語では、箇条記号と箇条本文のレイアウトに関して図 4 に示すような特徴をもっている。これらの特徴は、日本語や欧文の組版では見られない。これは、決してインデントについての意識が低いために、このような結果を生み出しているわけではない。組版について極めて高い規範意識に支配されていると思われる書籍や雑誌、あるいは公的な文書においても、このような組み方を見ることができる。この二つの言語の文書フォーマットでは、正統性をもったスタイルであると考えられる。

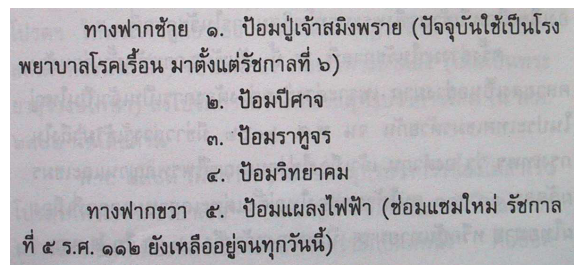


図 4: 例 1104

図 5 は、見出しの行のもっとも下のラインが、本文の最初の行のもっとも上のラインよりも下に入り込んでいる例である。

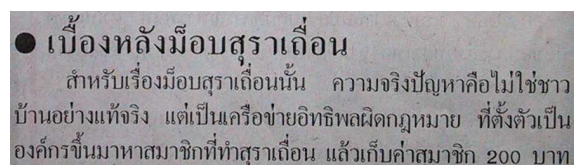
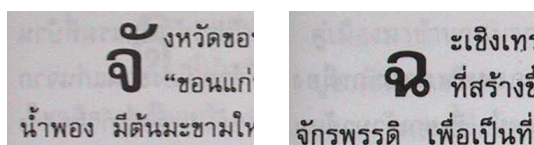


図 5: 例 0303

4.4 文字組み上の特徴

文字組み上の観点から興味深い事例は、欧文組版でいう 'drop cap' である。'drop cap' とは段落の先頭の一文字だけを取り立てて大きなサイズで組む方法である。タイ文字、ラオス文字などでは、基本となる子音文字の周囲に母音文字や声調記号、あるいは別の子音記号が付加

されることで一つの音節を表現する。これらの文字体系では、どの部分を取り出して、‘drop cap’を適用するかが問題になる。図6は、ともに‘drop cap’の例であるが、母音文字が伴う場合と伴わない場合のバランスが取れていない。また、ラオス語の例では、母音文字と基本となる子音記号とが分離したケースも見られている。

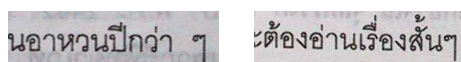


(a) 母音記号あり

(b) 母音記号なし

図6: 例1102、例1103

図7の二つの例は、どちらも最後に共通の記号が出現する。この記号は、前の音節をくり返すための記号であるが、この記号の組み方については、二つの例が示す通り、間隔をあける場合と間隔をあけない場合のふた通りがあるように観察される。この二つの事例が、規範による使い分けであるとすれば、スタイル指定はこのような使い分けを反映しなければならない。



(a) 間隔あり

(b) 間隔なし

図7: 例0802、例0803

5 結論とまとめ

本稿で事例として挙げたものは、筆者を含むDocSII委員会のメンバーにより注目すべき事例として取り上げられたものであるが、必ずし

もタイ文字、タイ語についての十分な知識を有するわけではない。

そのような制約はあるにしても、これらの事例が示すのは、文書フォーマットの要素は言語をまたがって共通の振る舞いをするわけではないということだろう。また、文書スタイルの標準化にあたっては、実際の書きことばの用例に基づいた要求抽出が必要であることも分かる。

このようなレイアウト、スタイル、組版に関わる処理の自動化は、人間可読文書と機械可読文書の両方の性格をあわせもつ電子化文書の自動処理（自動翻訳など）の場面で特に必要となる。また、近年の行政業務の電子化、あるいは「電子政府」の流れの中で、電子文書による情報交換の必要性は高まっている。

電子文書であっても、そのフォーマットは、当然に書きことばの伝統を十分に反映していなければならない。逆に、その言語の書きことばの伝統を十分に反映しない文書フォーマットは、その言語集団（speech community）から受け入れられにくい。言語集団の大半を構成する一般の話者は文書技術の専門家ではなく、文書技術が課す制約について十分認識していない。そのため、一般の話者は、電子文書という領域の外で行なわれている書きことば上の慣習を当然のこととして電子文書の中にも持ち込もうとする。

例えば、いわゆる「機種依存文字」の使用はその例である。オンライン・サービスなど初期のコンピューターネットワークの利用者の間では、丸付き数字など、機種依存文字の使用は避けるべきと考えられていたが、コンピューターネットワークが一般的になったときに、そのような考え方は大多数の利用者には支持されなかった。また、印刷される文書の作成にも、機種依存文字は普通に使用されている。それは、印刷された文書のあるべき姿について話者が抱

く規範意識や期待が、もともとこれらの文字の使用を想定しているということである。

他の例として、全角数字と半角数字の混在が挙げられる。商用印刷物の中には、ひと桁の数字に全角文字を使い、ふた桁以上の数字には半角文字を使用するという慣習に基づいて組まれているものが少なくない。そして、この慣習は、電子文書の中にも持ち込まれる傾向がある。

これらのことが示すのは、電子文書という新しい分野であっても、利用者が書きことばに対して既にもっている期待を反映しなければならないということだろう。もちろん、旧来の慣習のすべてが保存されるべきということにはならないが、電子文書が書きことばの重要な一側面として位置付けられるためにも、これまでの書きことばの慣習について検証する試みが必要であると思われる。

DocSII 委員会では、2003 年 9 月にモンゴル、また 11 月にスリランカにおける現地調査を予定している。モンゴル語、シンハラ語、タミル語における文書フォーマットの状況を現地の専門家と共同して調査し、これまでに筆者らが収集した事例以外にも、多言語組版を実現する上で考慮すべき文書フォーマット上の特徴があるかどうかを検証することが目的である。

謝辞

本稿は、CICC が実施した「多言語文書情報交換に関する国際規格共同開発調査」の一環として設置された「アジア文書情報交換委員会」の活動成果をもとにしている。筆者に委員として活動する機会を与えてくださった CICC に対し、また調査・分析に際して知見と助言を示してくださった小町主査に対して、この場を借りてお礼を申しあげる。

参考文献

- [1] Yannis Haralambous and John Plaice (2002)「組版・文書処理システムΩ」小林・安岡・戸村・三上編『インターネット時代の文字コード』（共立出版）所収
- [2] 小町 祐史（2003）「東南アジアの文書スタイルの現地調査結果」
<http://www.y-adagio.com/public/confs/docstyle/030519/0.htm>
- [3] 三上 喜貴（2002）「アジア圏諸言語に対応する文字コードの開発」小林・安岡・戸村・三上編『インターネット時代の文字コード』（共立出版）所収