UnicodeとXSLによる多言語組版

2003年12月29日 長野　花子（アンテナハウス株式会社　常務取締役）

要約 最初に、コンピュータによる多言語組版に取り組むにあたって問題になると考えられる点を一覧してみる。これらはいづれも単独でも難しい課題だが、技術が日進月歩で進歩しているので、それを理解して使いこなすのはさらに難しくなる。
Summary Let us first go over potential challenges in multilingual computer typesetting. Each of these items is already difficult enough on its own, and rapid progress in technology is making our mastery and utilization of such typesetting even harder.

組版対象のデータをどう作成するか？

コンピュータで文字情報を処理するには、まず、その文字情報が符号化されたデータとして作成されている必要がある。日本語のみを対象とするのと比べると、多言語のデータを作成するのは格段に難しくなる。(1)

1. 文字符号化方式の選定
   - 文字符号化方式（いわゆる文字コード）は主として国単位で、地域ローカルな文字コード表として標準化されてきた。しかし、地域ローカルな文字コード表でデータを表す方法では、多言語混在の文書を簡単に扱うことはできない。多言語編集、特に多言語混在の文書を編集、組版しようとしたらUnicodeが必須であろう。
   - Unicodeはどの言語まで使えるか？Unicodeの最新の標準化状況とそれを実装した製品にどのようなものがあるか？特にUnicodeは、かなり早いスピードで進化しているので最新の情報を正しく把握する必要がある。
   - Unicodeにはどのような問題があるだろうか？
   - 従来のシフトJISやASCII符号化によるテキストファイルでは使えるコードの種類が限られていた。これに対して、Unicodeのテキストでは、新しいコードがいろいろ定義されている。例えば、U+2000からU+200Fまでの16文字に次のようなコードがある。こういうコードは組版上どういう意味をもつのか？どうやって使いこなすのか？

U+2000からの16文字

```
2000;N # EN QUAD
2001;N # EM QUAD
2002;N # EN SPACE
2003;N # EM SPACE
2004;N # THREE-PER-EM SPACE
2005;N # FOUR-PER-EM SPACE
2006;N # SIX-PER-EM SPACE
2007;N # FIGURE SPACE
```

2. コンピュータの選定。ハードウエアとOSはどうやって選ぶ？
   - Macintosh、Windows2000/XP、Unix（Solaris等）、Linux、JAVA、どのような環境を選ぶか？
   - Windowsでは、Uniscribeという多言語処理層（ライブラリー）が整備されていて、これを使いこなすことでアジア圏の言語を含める多言語処理が可能になる。Internet ExplorerやMicrosoft WordはUniscribeを使って、多様なアジア圏の言語まで処理ができるようになっている。
   - 多言語処理ではWindowsがもっとも進んでいるが、ではJAVAではアジア圏の言語をどこまで処理できるだろうか？また、LinuxやSolarisなどのUNIXでの多言語組版の現状は？
3. データをコンピュータにどうやって入力するか。
   - データ入力用のソフトウエアにはどんなものがあるか？
   - キーボードの選定、キーボードをどうやって用意するか？キーボードは各国で標準化されていて、各国で販売しているPCには、その国の方式のキーボードが付随している。特に日

(1) この文書は、アンテナハウスの標準文書形式であるSimpleDoc.dtdに準拠したXMLで記述して、XSL Formatter V2.5で組版してPDF化したものである。この文書自体、Unicode、XML、XSLによる多言語組版の実例でもある。

本で販売しているPCに付属のキーボードで他の国の言語を入力するのはどうしたら良いか？

- IMEが必要か？その選定方法は？周知のように日本語を入力するには、ローマ字入力、かな漢字変換を行う方式が主流である。しかし、漢字の日本語読みを知らない外国人が、漢字を入力するのにローマ字読みで入力するのは無理ではないかと思う。それと同じで、中国語の入力方式として中国人にはピンイン入力が自然かもれないが、日本人がピンインで入力するのは難しいだろう。

4. データの表現法
   - アプリケーション依存のバイナリかアプリケーション独立のXML形式にするか？
   - 多言語処理を実現するにはXMLが一番良い。しかし、XMLを使いこなすのはハードルが高いのが事実である。XMLのタグはそれほど難しいものではないが、一般の人達はタグを異様に恐れる傾向がある。XMLへのハードルをどうやって下げたら良いか？
   - XMLを使う場合、データ構造（スキーマ）を設計しなければならない。
   - 新しくデータ構造を定義するのではなく、既存のDTD/スキーマを使えないか。例えば、DocBook.dtdは使えないのか？
   - 新しい標準DTD/Schema定義の動向は？
5. 編集ソフトウエアの選定
   - 使い慣れた編集ソフトを使えるかどうかは、文書作成の生産性に非常に大きな影響を与えるので、どのような編集ソフトを使うことができるかを選択することは非常に重要である。この観点からは、Microsoft Wordを使えれば非常に便利だ。Microsoft Wordは多言語編集ソフトとしてどこまで使えるか？
   - 多言語対応を標榜している編集ソフトは多数ある。しかし、例えば、英語、西欧の言語、日本語、中国語、韓国語、アラビア語、ヘブライ語、タイ語をひとつのバージョンですべて編集できるソフトは数が少ない。もし、言語別に編集ソフトを切り替えなければならないなら、多言語が混在する文書は作成できない。また、言語別に切り替えるとなると、操作を新しく覚えたり、データの互換性の問題も出るので、これは避けたい。しかし、どうしても言語別に編集ソフトの切り替えが必要になったらどうするか？
   - 多言語をWYSIWYG、あるいはWYSIWYGに近い形で編集できるツールがあるか？あるとして、どのソフトを選択するのが最適か？
   - XMLでデータを作成するには、スキーマに従ったデータの入力編集作業を支援するためのツールが必要である。そのようなソフトがあるか？
   - 専門家が文書を作成する場合は、訓練、学習ができるので、タグを見せるタイプのXML編集ソフトでも使用できる。多言語でそのようなことのできるXML編集ツールがあるか？あるとして、どのソフトを選択するのが最適か？

組版の方法

1. レイアウトを頻繁に変えることができ、かつ、WYSIWYGで編集でき、レイアウト編集した結果をXMLのソースデータにリアルタイムで反映できるような、本格的な多言語XML組版ソフトは存在しないのか？なぜ存在しないのか？
2. 文字を画面なり、紙、PDFなりに視覚化して表すには、フォントが必須である。では、Unicode対応フォントにはどのようなものがあるか？
3. PDFを作成して配布したり、印刷しようとすると、フォントのアウトラインの埋め込みが必須である。従って、多言語組版で使用するフォントは、アウトラインの埋め込みが許可されたフォントでないとだめだろう。多言語の組版をしようとするとき、使えるフォントにはどんなフォントがあるか？
4. XSL-FOによるレイアウト指定方法
   - XSL-FOはどこまで使えるか？どの程度まで、複雑なレイアウトが指定できるか？
   - XSL Formatterは、XSL仕様を満たすXMLの多言語組版ソフトであるが、どんな特徴があるか？
   - 組版規則が言語別に違うケースにも対処できているのか？
   - 文章の中に異なる組版規則の言語が混在する時にも対処できるのか？
   - 右から左へ記述する言語と左から右へ記述する言語が混在しても対処できるのか？

印刷・PDF作成方法

1. 多言語の印刷はどうやってやるか？
2. 印刷用PDFとWeb用PDFの違い、使い分け。

その他

1. 目次の作成方法や索引の作成方法
2. 索引のソート順、言語別のソート規則、多言語混在時のソート規則

多言語組版の基礎知識

文字と言語

言語は文字によって記述される。言語をコンピュータで扱うためには、まずその前提として、言語を表記する文字を扱うことができなければならない。コンピュータで文字を扱う時は、普通は文字の種類を集合として規定し、番号付けした符号化文字集合を使う。従来は、各国・地域別に規定された符号化文字集合が使われてきた。次の表は主要な言語が主にどのような地域別の文字コードで表されるかを示す。

Unicodeの歴史

1991年10月	Unicode 1.0.0版発行
1996年 7月	Unicode 2.0.0版発行
1999年 9月	Unicode 3.0.0版発行
2002年 3月	Unicode 3.2.0版発行
2003年 4月	Unicode 4.0.0版発行

Unicodeは単に符号化文字集合を定義するだけではなく、テキスト処理のための各種の基準を定義している。①文字を書き進める方向を規定するUnicode文字データベース、②禁則文字など文字の改行特性を規定するUnicode Line Breaking Properties、③双方向性処理を規定するUnicodeBIDI、などの技術レポートはUnicodeの仕様の一部である。これらのレポートは、完全とは言えないこともあるが、アプリケーション・プログラムを作成する際に参照として貴重な情報源である。Unicodeの仕様とこれらの技術レポートによって、多言語組版エンジンの開発が非常に容易になった。

OSとアプリケーションの内部文字コード

1980年代から1990年代までのパーソナル・コンピュータのOSは各地域別の文字コード表に基づいていたため、OSの上で動作するアプリケーション・プログラムも地域別であった。例えば、日本語Windows98/Meは、OS内部ではシフトJIS（JISX0201とJIS X0208）でテキストと文字処理を行っている。このためWindowsMeで動くアプリケーションでは、A with diacresis : Ä、O with diacresis : Ö、U with diacresis : Üなどのラテンの特殊文字を正しく処理するのが容易ではない。これらの文字は日本語の半角カタカナと文字コードがぶつかってしまうためである。

MicrosoftのWindows2000/XPは、OS内部の処理がUnicodeが基本になっていて、多言語処理機能が飛躍的に強化されている。MicrosoftのOSで多言語処理を行おうとしたら、Windows2000またはXPを選定する必要がある。

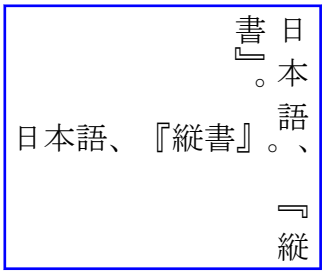
アプリケーション・ソフトウエアには、内部的にUnicodeでデータを処理しているものと、ローカル文字コードで処理しているものがある。多言語の処理を行おうとしたら、内部的にUnicode処理をしているアプリケーション・ソフトウエアを選択する必要がある。例えば、XSL FormatterやMicorosoft Word2000/XPは、Unicodeアプリケーションであるが、FrameMakerは、Unicodeアプリケーションではない。

アプリケーションの役割

アプリケーション・ソフトウエアがUnicodeを扱えるだけでは多言語処理ができるとは言えない。Unicodeと多言語処理の間には、超えなければならない問題が沢山ある。文字コードのレベルでいうと次のようなものである。

グリフ置換

日本語や中国語（繁体字）では同一文章を横書きと縦書きができる。横書きと縦書きでは句読点や括弧類は同じ文字コードを異なるグリフで表示・印刷する必要がある。組版エンジンは自動的にグリフの入れ替えを行わねばならない。



また、アラビア語では、同じ文字コードであっても単独の時、単語の開始位置に表れる時、中間、終了位置に表れる時でグリフが変化する。アラビア語を正しく扱うソフトウエアは文字の位置によるグリフの入れ替えを実現しなければならない。