文字画像化を用いた多言語表示手法の提案

朱　槿　　浦野　義頼
早稲田大学大学院　国際情報通信研究科　〒367-0035 埼玉県本庄市西富田大久保山 1011

あらまし　近年、情報通信技術の高速な発展と、モバイル通信設備の普及拡大に伴う国際化の流れにより、携帯端末における多言語対応は求められるようになった。ただし、世界中で使われている文字の数は数万があるが、携帯端末はモバイル設備として計算能力と機種差異の制限もあるために携帯端末での多言語対応は依然として困難である。キャリア・機種に関わらず、世界各国の文字を携帯端末で表示させるために、本研究では各種文字そのものの構造特徴を分析してコンポーネントの組み合わせ方式を利用した、書字方向が自由でかつ複数の言語が混在できる効率的な多言語表示手法について提案してきた。本稿ではその概要及び実験システムの評価検討結果について述べる。

キーワード　モバイル、多言語、表示方法、自然言語、データベース、XML

Proposal of Multilingual Display Method Based on Image Conversion of Character & Component

Jin　Zhu　　Yoshiyori　Urano
Graduate School of Global Information and Telecommunication Studies, Waseda University
1011 Okuboyama Nishi-Tomida Honjo-shi Saitama 367-0035 Japan

Abstract　Recent years, the world is surely shrinking as information communication and computation advances proceed at a breath-taking pace. Under this background, the demand of portable electronic device which can support multilingual is growing day by day. But, there are about several ten thousand characters in all of the writing systems of the world and the computation ability portable electronic device is limited, and there are some big discriminations of portable electronic device's hardware in different career. Therefore, the multilingual support of portable electronic device is extremely difficult to realize. In this paper, we have analyzed character structure of most of the writing systems, and propose a multilingual display method which using components' combination. Moreover, we show results of simulation experiments. The results demonstrate that using this method, all of the languages in the world can be displayed on different types of mobile, and our method can reduce the cost of data translation efficiency comparing with the usual method.

Keyword　Mobile, Multilingual, Display Method, Natural Language, Database, XML

1　はじめに

近年、情報通信技術の革新や自由貿易体制の拡大に伴い，経済活動をはじめあらゆる側面で世界の一体化が急速に進んでいる。このような背景に

おいて、外国語を勉強する人数が大幅に増加し、世界範囲の人口流動も頻繁になっている。外国語学習の支援、及び外国人住民の日常生活支援と緊急災害時支援を行うために、多言語による情報の提供は不可欠である。携帯電話は「時間や場所を選ばず情報を入手できる」という特徴があるので、語学学習者や外国人住民への情報提供は携帯端末を用いることが最も有効であると考えられる。

日本の携帯端末においては、全ての端末で一般的に用いられる文字コードはシフト JIS であるため、日本語とアルファベット以外の文字は基本的にそのままでは表示できず、そのための変換処理が必要になる。従来の研究では，文字変換表方法と文章画像化方法があるが、前者の方式については、一部の言語（主に漢字を利用する言語）でしか適用できない制限がある。後者の方式では変換した画像ファイルが、携帯端末のキャリア・機種によって機能、画面サイズ、表示形式、データ容量等が異なるため、ユーザが使用する端末機種の判明，及び画像サイズの適当の変換をしなければ送信した画像が表示されない場合があるという問題がある。

上記の問題点と制限を解決し、携帯端末のキャリア・機種に関わらない、かつ効率的な多言語表示を実現するために、本研究では世界の文字の構造を分析したうえで、コンポーネント（component）の組み合わせ方式を利用した、携帯端末における複数の言語が混在できる多言語表示手法を提案する。

2　文字の分類

2.1　文字の構造的分類

現在、世界各種の文字を分類するための様々な分類基準が存在する。伝統的によく用いられる分類は「表音文字/表意文字」による分類であるが、しかし、音と意味は、文字の構造を示してはいない。即ち、文字の表示処理に関しては、音価ではなく、文字の持つ構造と処理単位で分析する必要がある。世界の文字をその構造で分類すると、字形がひとつのまとまりになっており、構造的にそれ以上分解することができない単体構造の「単体文字」といくつかの小さな構成単位が組み合わせさってきた複合構造の文字である「合体文字」に分類される。

ラテン文字をはじめほとんどのアルファベットを使っている文字は「単体文字」であり（一部のアルファベット言語の表示について、母音字にはアクセント符号を使用するが、母音字とアクセントの組合せ方式は非常に単純であるために、本研究ではこのような文字も単体文字として扱う）、中国語、日本語などで使われる漢字と朝鮮語で使われるハングルは「合体文字」に属する。

2.2　合体文字におけるコンポーネントの構造と数

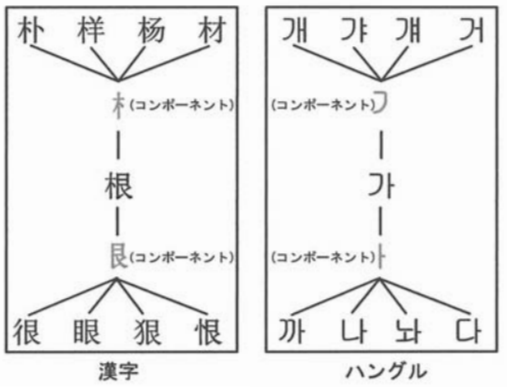


図１　合成字におけるコンポーネント構造

合体文字そのものの構造から考えると、図１の左側のように漢字の体系は、表意・表音の偏旁と表意でも表音でもない記号の組み合わせによって構成されている。図１の右側に示したようにハングルでは、子音・母音を表すそれぞれの字母があり、それぞれの字母が１つの音を表すアルファベット型の文字である。しかしながら、ハングルの字母はローマ字のように一列に並べて書かれるのではなく、基本的に１音節ごとに一まとめにして１文字を成す。このような合体文字の構成単位をコンポーネントという。コンポーネントは漢字とハングルを形作る最も基本的な構成要素であり、数少ないコンポーネントの組み合わせによ

り、数万の文字が作成できる。

構造方法	構造図形	字例
単体字構造		个
左右構造		请　树
上下構造		尘　鼻
包む構造		庙　司　还 闪　巨　凶 因

表 1　漢字の構造方法

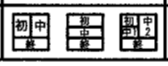
構造方法	構造図形	字例
子音＋母音 (初＋中)		가　드　과
子音＋母音＋子音 (初＋中＋終)		간　읔　된

表 2　ハングルの構造方法

ハングルでは 19 個の子音字母（そのうち基本字母が 14 個、合成字母が 5 個）と 21 個の母音字母（そのうち基本字母が 10 個、合成字母が 11 個）があり、ハングル文字を構成するコンポーネントの総数は 40 個である。

漢字を切り分けるときは，段階的に切り分けていく方法をとる。つまり、一度に多くの部件に切り分けるのではなく、何段階かに分けて切り分けるのである（図 2)。そうすれば，はっきりとその字の構成がわかる。文字の分解段階によって、コンポーネント数ははるかに違う。本研究では、効率的な表示を実現するために参考文献[6]のような漢字を 7 つの分解段階のうちの第 2 段階まで切り分け、コンポーネントの総数は 1149 個である。

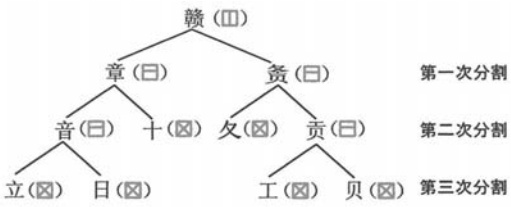


図 2　漢字の切り分け

コンポーネント構造とは，合体文字におけるコンポーネントの互いの位置関係を図示化したものである。コンポーネント同士の構造方式には、漢字とハングルそれぞれ、表 1 と表 2 のようなものがある。

文書の携帯端末での表示においてコンポーネント構造という概念を採用することにより，50,000 字を超えている漢字（参考文献[12]）と 11,172 字のあるハングル（参考文献[1]）ごとの大量処理から，文字を構成する数少ないコンポーネント及びコンポーネントの組合せへの少量処理になり，本研究の基礎となる.

2.3　書字方向による分類

書字方向	言語種別
右横書き	アラビア語、ヘブライ語
左横書き	欧米諸語
右縦書き	日本語、中国語
左縦書き	モンゴル語

表 3　書字方向による分類

表 3 の分類のように、世界に存在する文書は、その言語および表記する文字体系の組み合わせによって文字を書き進める方向（書字方向）が異なり、大きく分けて文字を縦に連ねる縦書きのものと、左右に並べる横書きのものが存在する。縦書きの中には行を右から左へ進めるものとその逆があり、また横書きでは行中の書字方向が左右それぞれから進めるものがある。

3　提案手法

3.1　概要

本研究の提案では、前述の文字のコンポーネント構造方式の特徴を利用し、中間サーバで文章を構成する文字の語種と書字方向および文字に含まれたコンポーネントを分析する。そして、分析結果により、異なるコンポーネントごとに画像ファイルを作成し、さらに文字の書字方向とコンポーネント組み合わせ方式を記述した XML 文書を作成する。一方、携帯端末で XML 文書による文

字・コンポーネントの並び順、大きさを確定し、コンポーネント画像ファイルを活字のように組み合わせて文章を再作成する。この処理概要を図3に示す。

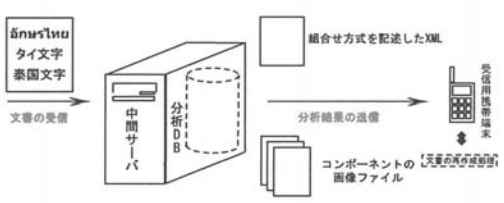


図3　システムの全体像

3.2　中間サーバにおける処理

図4に、中間サーバにおける処理を示す。中間サーバでは主に「①言語種別と書字方向の判断」「②文字構造と文字構成用コンポーネントの分析」「③コンポーネント画像ファイルの作成」「④組み合わせ方式XML文の作成」が行われる。以下に、これらの処理について述べる。

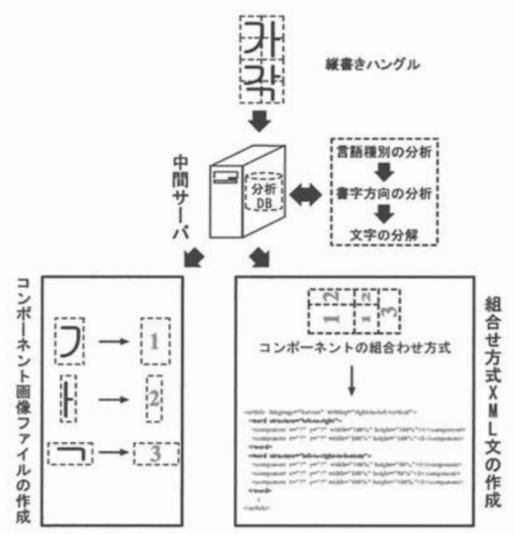


図4　中間サーバにおける処理

①言語種別と書字方向の判断

言語は種別によって文字の構造と書字方向が異なり、文字を表示するためにはまずその文字の言語種別を弁明しなければならない。そこで、中間サーバではまず分析DBを利用して、文字の種別と書字方向を分析して判別する。

②文字構造と文字構成用コンポーネントの分析

文字の種別と書字方向を判明した後、活字として扱うコンポーネント画像ファイルの作成のために、文字を構成するコンポーネントとそのコンポーネントの組み合わせ方式を分析する必要があるので、文字の分解基準を基づいて，一つずつ文字の構造特徴，構成コンポーネントを分析し，構造分類別とコンポーネント分類別に分類して文字分析DBを用意する．文書を受信した際，文字分析DBを利用して文章を構成する文字の並び順，構造方式，構成コンポーネント，さらに文字におけるコンポーネント位置関係を分析する．

③コンポーネント画像ファイルの作成

前述のように、文字は構造的分類による「単体文字」と「合体文字」に分れる。「単体文字」はアルファベットであり、大半は20から30の基本的な記号で構成され、最も少ないのがソロモン諸島のロトカス語の11文字、最も多いのがカンボディアのクメール文字の74字である。このような文字を携帯端末で表示させるために、本研究では文章に現れた各アルファベットを活字のように画像ファイルに変換したうえ、携帯端末で書字方向のXML文による画像ファイルを組み合わせて文章を再作成する方法を採用している。「単体文字」の作成アルゴリズムを図5に示す。

```
単体文字のコンポーネント画像化アルゴリズム
ComponentGraphicMaker_SingleChar (c,R)

入力：分析DBで文書に含まれた文字を順番で分解した文字
      （コンポーネント）のリストc(m),
      mは文章に含まれた文字の総字数,
出力：文章に含まれた異なるコンポーネントを蓄えた領域R
1. for (i←1 to m)
2.     if  Rに蓄えた文字を調べて，c (i)と同じ文字が
           Rの内部にあれば
3.         then  次の文字の判断にいく
4.     else   c (i)の文字をRに蓄える
5.     end if
6. next
7. return R
```

図5　単体文字のコンポーネント画像化アルゴリズム

携帯端末で「合体文字」を再作成するときの基本要素として、文字を構成するコンポーネント画像ファイルの作成が必要である。コンポーネント画像ファイルを作成するために、図 6 のようなアルゴリズムを定義した。

合体文字のコンポーネント画像化アルゴリズム
ComponentGraphicMaker_CombinationChar (c,R)

入力：分析DBで文書に含まれた文字を順番で分解した構成
　　コンポーネントのリストc(m,n_i),
　　mは文章に含まれた文字の総字数,
　　n_iは文字iを構成するコンポーネントの総数。
出力：文章に含まれた異なるコンポーネントを蓄えた領域R

```
1. for (i←1 to m)
2.     for (j←1 to ni)
3.         if  Rに蓄えた部件を調べて，c (i,j) と同じ部件が
               Rの内部にあれば
4.             then  次の部件の判断にいく
5.         else   c (i,j) の部件をRに蓄える
6.         end if
7.     next
8. next
9. return R
```

図 6　合体文字のコンポーネント画像化アルゴリズム

④組み合わせ方式 XML 文の作成

携帯端末でコンポーネント画像ファイルを組み合わせて，文書を再作成するために，文字分析 DB を利用して文書を構成した文字，文字を構成したコンポーネントの並び順、及び書字方向を分析する必要がある．分析の結果で XML 文を作成し，携帯端末に送信する．作成した XML 文書の例は図 7 の通りである。

```
<article language="korean" writing="right-to-left vertical">
  <word structure="left-to-right">
    <component x="?" y="?" width="100%" height="100%">1</component>
    <component x="?" y="?" width="100%" height="100%">2</component>
  </word>
  <word structure="left-to-right-to-bottom">
    <component x="?" y="?" width="100%" height="50%">1</component>
    <component x="?" y="?" width="100%" height="50%">2</component>
    <component x="?" y="?" width="100%" height="100%">3</component>
  </word>
    :
</article>
```

図 7　XML の構造

3.3　携帯端末における処理

①左横書き携帯端末における縦書き表示と右横書き表示

文字によっては、物理的な横書き表記方式もしくは縦書き表記方式があるので、多言語の文字を表示するために、縦書きでの上下双方の表示と、横書きでの左右両方向の表示が可能でなければならない。ただし、左横書きの英語を用いる米国で発展した携帯電話は、ほとんど左横書きの文字表示方式を採用している。また携帯端末としての計算能力などの制限があるので、右横書きと縦書き表示に対する双方の対応ができないことが現状である。本研究では、図 8 のように中間サーバで作成された XML 文書により、コンポーネント画像ファイルの回転と文字の順番の再割り当てを行い、左横書き携帯端末で文字の縦書き表示を実現する。右横書きについては一つ一つの文字の鏡像をとって段落を作成した後、全体の鏡像をとる方法を利用して実現する（図 9）。

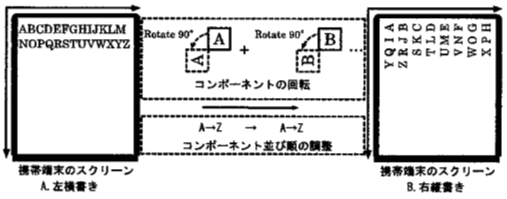


図 8　縦書きの実現

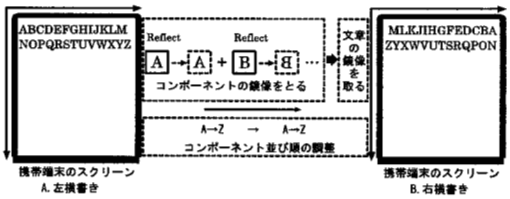


図 9　右横書きの実現

②XML 文による文書の再作成

携帯端末側では、中間サーバから送信してきた画像ファイルと XML 文を利用して組合せプログラムで文書を再作成することができる。再作成のモデルは図 10 のようである。

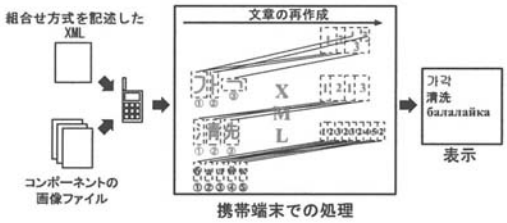


図 10　携帯端末における処理

4　評価

提案手法の機能・性能を検証するために評価実験を行った。(図 11, 12, 13)

以下に、試作・評価環境を示す

中間サーバ側：

OS: Fedora Core 5

Server: Apache 2.0.59

DB: Mysql 5.0.26

SP: PHP 4.4.4

XML：1.1

携帯端末シミュレータ：

AU：Openwave SDK 6.2K

Vodafon：ウェブコンテンツヴューア 5.0.1.124

Docomo: i-mode HTML Simulator II

図 11　多言語表示の実現

図 12　縦書きと右横書きの実現

図 13　異なるキャリアにおける表示

測定結果により、本手法と従来の手法の機能別

比較した。その結果を表 4 に示す。

	提案手法	文字変換表方法	文章画像化方法
多言語対応	可	不可	可
キャリア・機種の制限	無	無	有
右横書き・縦書き対応	可	不可	可

表 4　提案手法と従来手法の機能比較

また、性能について、従来の「文章画像化」手法と本研究で提案した手法を用いて、同じシミュレーション環境で実験を行った。図 14 のように、従来手法を利用した場合、文字数の増加に対し、送信するデータサイズは線形増加であるが、提案手法を利用した場合は、文字を構成するコンポーネントの数が有限であるので、表示する文字数がいくらであっても、送信するデータサイズは全てのコンポーネント画像ファイルの総サイズを超えない。つまり、大量文字を表示する場合、提案手法は従来の手法より顕著に効率的といえる。

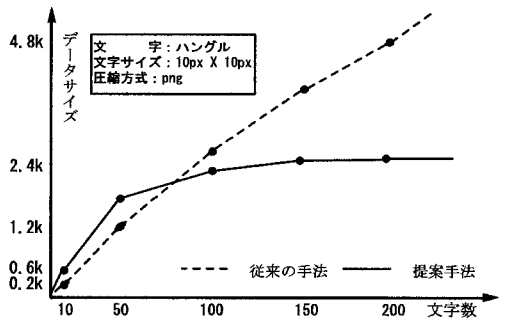


図 14　提案手法と従来手法の性能比較

5　おわりに

本研究では、文字そのものの構造的と特徴に基づいて、キャリア・機種に関わらない携帯端末における多言語表示手法を提案した。さらに、実験用アプリケーションを構築し、本手法を評価するために実験を行った。その結果、本手法は従来の方法の機種依頼性などのデメリットが解決できることを証明した一方、大量文字を表示するとき、従来の方法より顕著に効率であることも確認できた。

参考文献

[1] 市河三喜，高津春繁，服部四郎 “世界言語概説” 研究社，2000

[2] Andrew Robinson “The Story of Writing” Thames & Hudson; New Ed edition, 1999

[3] 池田 紘一，今西 祐一郎 “文字をよむ” 九州大学出版会，2002

[4] 矢島 文夫“文字学のたのしみ”大修館書店，1977

[5] 樊静 “漢字信息字典”上海科学出版社，1988

[6] 張静賢 “漢字教程”北京語言学院出版社，2002

[7] 蘇培成 “二十一世紀的現代漢字研究”書海出版社，2001

[8] 尹斌庸，John S. Rohsenow “現代漢字”華語教学出版社，1994

[9] 王寧 “漢字構形学講座”上海教育出版社，2004

[10] 奥村彰三，前田正弘 “漢字画像から文字要素の自動抽出”情報処理学会論文誌，Vol. 32, No. 1, pp. 50－61, 1991

[11] 西野 嘉章 “歴史の文字 記載・活字・活版”東京大学出版会，1996

[12] 徐中舒“漢字大字典” 四川辞書出版　湖北辞書出版社，1986

[13] Mark Davis “The Bidirectional Algorithm”, http://www.unicode.org/reports/tr9/