E_015

漢字の構造方式を利用した中国語表示方式の提案

Proposal of Chinese Display Method Using Composition Method of Chinese Character

朱 槿† 浦野 義頼†
Jin Zhu Yoshiyori Urano

1. はじめに

近年，IT技術の高速発展に伴い，世界経済が一体化に進んでいる．その結果，国境と言語のギャップを越える情報交換へのニーズが多くなっている．このようなニーズを対応するため，単一端末（PC，携帯端末など）での多国語対応への要求が高まっていき，世界各国の言語に対応できる多国語技術（Multilingual technology）実現の研究は最も重要な研究テーマの一つになっている．PCでは，CPUの計算処理能力の大幅増強，大容量記憶装置の出現，各国政府や研究機関の自国言語の文字コード，フォント作成及び入力方法と表示方式への研究開発に対する支援と推進，UNICODEの普及などに従い，多国語対応問題がほとんど解決された．しかし携帯端末はPCと比べて計算能力と記憶容量が低いといった，ハード面の制限がある．また，携帯電話の機種多様性により，異なる通信サービス事業者と端末メーカーが提供する携帯端末間の互換性が低いという問題点も現実的にある．これらの問題により，携帯端末での多国語対応は依然として困難である．

日本の携帯端末においては，日本語と中国語の文字及び文字コードの差異の問題より，日本語の携帯端末での中国語の入力，表示は依然として困難である．従来の研究では，中国語漢字と日本語漢字の変換表を使って中国語漢字を日本語漢字に変換する方式とメール，ニュースなどの中国語文章を画像ファイルに変換する変換システムを介し，変換した画像ファイルを携帯端末に送信して表示させる方式がある．ただし，前者の方式では，変換表にない漢字で偏（ヘン）や旁（ツクリ）を組み合わせて表現できるものは漢字一文字を２つの文字の組み合わせで表し，それでも表現できない漢字はピンイン（中国語の表音文字）で表わす方式になっているため，変換された結果は完全な中国語とはいえない．後者の方式で変換した画像ファイルは，携帯端末の機種によっては，機能，性能，解像度などが異なるため，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．

したがって，本研究では機種間互換性のボトルネックを超えて大多数の携帯端末で純粋な中国語を完全に表示させるために中国語漢字の構造方式，活字印刷の原理に基づき，上記の問題を解決することを目的とする．送信者と受信者の間に中間サーバを設置し，送信者から送ってきた中国語文章を中間サーバで文字ごとに分割し，さらに部件に切り分ける．そして，分析した結果により，文章に含まれた異なる各部件を活字のような画像ファイル及び部件組合せ方式を記述したXML文書を作成し，携帯端末に送信して文章を再作成する方式を提案する．

2. 中国語漢字の構造方式

2.1 単体字と合成字

漢字はその構造から，単体字（独体字）と合成字（合体字）の二つに分類することができる．筆画をそのまま組み合わせてできた字を単体字という．単体字は字形が一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漢字である．たとえば，“人・口・山”などである．一方，いくつかの小さな構成単位から構成された字を合成字という．合成字はみな字形が分解でき，いくつかの部分に分けることができる複合構造の漢字である．たとえば，“江，知，吉”などである．現代中国語で用いる漢字は，大多数が合成字で単体字は少ない．

2.2 部件（コンポーネント，component）

合成字の構成単位を部件という．部件は漢字を形作る最も基本的な構成要素であり，漢字の字形は，「筆画」「部件」，「整字」の三つのレベルに分けることができる．30あまりの筆画を組合せて600～700の「部件」をつくり，その600～700の部件を組合せて数万の漢字ができる．それを図に示すと図1のようになる．

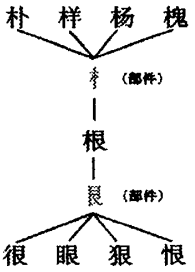

図1 漢字と部件の組合せ

このように，部件は漢字を形づくる最も基本的な構成要素であることがわかる．多くの漢字は，二つ以上の部件によって成り立っている．参考文献[1]では，7785字の漢字の部件の数について統計をとっている（表1）．

類別	字例	字数	割合（%）
一つの部件からなる字	人	323	4.149
二つの部件からなる字	都	2650	34.040
三つの部件からなる字	煤	3139	40.321
四つの部件からなる字	整	1276	16.391
五つの部件からなる字	矗	323	4.149
五つの部件以上の字	整	75	0.950

表1 漢字を構成する部件数の統計結果

文書の携帯端末での表示において漢字の部件という概念を採用することにより，数多くの漢字ごとの大量処理から，数少ない文字を構成する部件及び部件の組合せへの少量処理になり，本研究の基礎となる．

2.3 部件に切り分ける方法

†早稲田大学国際情報通信研究科

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．そうすれば，はっきりとその字の構成がわかる．たとえば（図2）：

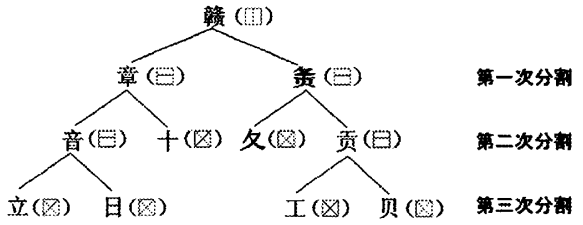

図2　漢字の切り分け

1983年1月から1984年5月まで，『辞海』（1979年版）に収められる16296字と，『辞海』では未収だがGB 2312 - 80国家標準「信息交換用漢字編碼字符集・基本集」に収められる43字の，計16339字について，中国文字改革委員会と武漢大学が共同して，コンピュータによる部件の分析を行った．そのなかの11834字の簡化字と未簡化字について段階ごとに行った部件の一覧表は，表2のとおりである．

分割の段階	部件総数
第一段階	2556
第二段階	1149
第三段階	477
第四段階	168
第五段階	39
第六段階	9
第七段階	2

表2　分割の各段階の部件総数統計

2.4 部件構造

部件構造とは，漢字における部件の互いの位置関係を図示化したものである．参考文献[1]に収められている7785個の正体字のうち，一個の部件で構成されるのはわずかに323字であり，総数の4.149%に過ぎない．このように，95%以上の漢字については，部件のパターンの組合せとなっている．

漢字は，いくつかの部件をそれぞれ四角い箱のなかにおさめている．長期にわたる書写の実践が基礎となり，漢字に数種類の基本的な部件構造のパターンがつくられた．主なものは表3のようなものがある．

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　齿 困

表3　部件構造

3. 実現

3.1 基本方針

本研究の提案では，部件の組み合わせである中国語漢字構造方式の特徴を利用し，中国語文章を構成する漢字を分析して，漢字に含まれた部件を抽出する．そして，抽出した異なる各部件を活字のような画像ファイル及び部件組合せ方式を記述したXML文書を作成し，携帯端末でXML文書による部件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．

3.2 漢字の分解基準

携帯端末の記憶容量及びCPUの計算能力には限界があるため，巨大な画像ファイル，或いは漢字の過度の分解は結果として携帯端末への負担を招く．そのため，本研究においては，漢字の分解基準として以下の前提条件を設けた．

(1) 左右構造と上下構造の漢字だけを分解する．

前述のように，漢字の構造方式は左右構造，上下構造などがある．左右構造と上下構造の漢字は部件の単純な横方向と縦方向の組み合わせで作成できる．

また，統計分析によれば，左右構造の字は現代漢字の大多数を占めている．参考文献[1]の統計では，左右構造の字は5055個で，収められた全字数7785字の64.933%を占める．これに次ぐのが上下構造の字で，全部で1643個，7785字中の21.105%である．つまり，左右構造と上下構造の漢字は，7785字中の86.038%を占める．

効率的なシステムを構築するために，本研究では，左右構造と上下構造の漢字だけを部件に分解する．その他の構造の漢字は分解しない．

(2) 漢字を第二段階まで切り分ける．

漢字を切り分ける際，細かい部件まで分解した場合，中間サーバでの分解処理と画像ファイル，XMLを作成するための作業量，及び携帯端末で文章を再描画するときの処理量が大量になり，結果としてレスポンスの悪化につながる．そのため，本研究では，漢字を第二段階まで切り分ける．

3.3 システムの構成

以上の基準にしたがって，使われる中国語漢字の中で使用頻度の上位15個の部件を含む左右構造の漢字15個，上下構造の漢字15個，及び本研究で分解しないその他の構造の漢字15個を分析できるようなシステム（図3）の試作を行った．漢字の構造方式を利用した中国語表示実験システムは，中間サーバ側のサブシステムと携帯端末側のサブシステムの2部分で構成されている．

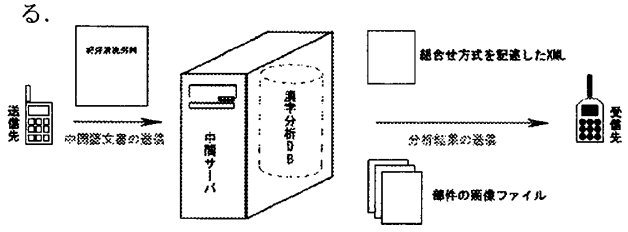

図3　システムの全体像

(1) 中間サーバ側のサブシステム

図４のように、中間サーバ側のサブシステムでは，漢字分析 DB を利用して中国語文書の漢字の並び順を追って構成部件を分析する．続いて，分析結果により，分解できる漢字の異なる構成部件ごと及び分解できない漢字ごとに画像ファイルを作成する．さらに，文章における漢字の並び順と部首の組み合わせ方式を記述する XML 文を作成する．

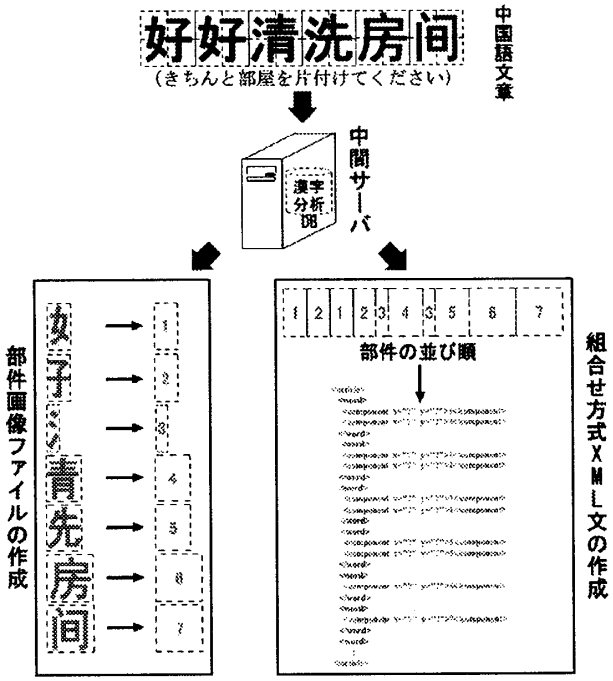

図４　中間サーバ側のサブシステム

文書の分析と漢字分析 DB

上記の漢字の分解基準を基づいて，一つずつの中国語漢字の構造特徴，構成部件を分析し，構造分類別と部件分類別に分類して漢字分析 DB を用意する．送信者からの中国語文書を受信した際，漢字分析 DB を利用して中国語文章を構成する漢字の並び順，構造方式，構成部件，さらに漢字における部件位置関係を分析する．

部件画像ファイルの作成

携帯端末で漢字を再作成するときの基本要素として，漢字を構成する部件画像ファイルの作成が必要である．部件画像ファイルを作成するために，図５のようなアルゴリズムを定義した．

```
部件画像ファイルの作成アルゴリズム
ComponentGraphicMaker (c, R)

入力：漢字分析DBで中国語文書に含まれた漢字を順番で分解した
      構成部件のリストc (m,n_i)，mは文章に含まれた漢字の総
      字数、n_iは漢字を構成する部件総数。
出力：中国語文書に含まれた異なる部品を蓄えた領域R
1. for (i←1 to m)
2.     for (j←1 to n_i)
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５　部件画像ファイルの作成アルゴリズム

組み合わせ方式 XML 文の作成

携帯端末で部件画像ファイルを組み合わせて，中国語文書を再作成するために，漢字分析 DB を利用して中国語文書を構成した漢字，さらに漢字を構成した部件の並び順を分析する．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６の通りである．

```
<article>
    <word>
        <component x="?" y="?">1</component>
        <component x="?" y="?">2</component>
    </word>
    :
</article>
```

図６　XML の構造

(2) 携帯端末側のサブシステム

携帯端末側のサブシステムでは，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，組合せプログラムで中国語文書を再作成することができる．再作成のモデルは図７のようである．

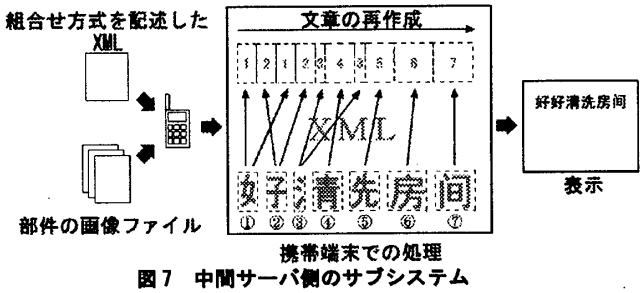

図７　中間サーバ側のサブシステム

4. 今後の課題

漢字構造分析ツールの開発

本研究で試作したサンプルシステムでは，部件構造の特徴を表している特殊的な漢字を 45 個しか作らなかったが，古代から現在まで，漢字の総数は全部で五万字を超えている．単に現代中国語で使われている漢字だけでも，およそ一万存在する．この大量の漢字を一つずつ人力で分析して漢字分析 DB を作成するのは大変である．低コストで漢字分析 DB を構築できるように，筆者らは漢字の部件構造という特徴と OCR 技術を基づいて，漢字構造分析ツールの開発を行いたい．

5. 終わりに

本研究では，一部の漢字を部件に分解して，部件の画像ファイルと組合せ XML を作成し，携帯端末で組合せプログラムによって再合成を確認することができた。今後においては，全ての漢字を表現できるよう，さらなる努力が必要である．

参考文献

[1] 樊静.“漢字信息字典” 上海科学出版社，1988
[2] 張静賢.“漢字教程” 北京語言学院出版社，2002
[3] 蘇培成.“二十一世紀的現代漢字研究” 書海出版社，2001
[4] 尹斌庸，John S.Rohsenow.“現代漢字” 華語教学出版社，1994

[5] 王寧.“漢字構形学講座”上海教育出版社，2004
[6] 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