人文科学とコンピュータ　42－10
（１９９９．５．２１）

e漢字の開発と利用に関連する諸問題

丹　羽　正　之

京都大学人文科学研究所

多数の漢字フォントを効率よく作成するために，まず漢字を部品単位に分解し，漢字の構造や部品の種類，変形などを調査した．それらのデータに基づき，部品から漢字を合成することで，５万字を超える多数の漢字を迅速に開発することができた．本論文では，多数の漢字に対して，分解と合成の手法が，フォント作成にきわめて有効であることを示すとともに，漢字の意味的，音的な分類にも役立つことを示した．

Development and Usage of eKanji - over 50000 Kanji characters

Masayuki Niwa

Institute for Research in Humanities, Kyoto University

In order to make over 50000 Kanji characters rapidly and less costly, we have researched the structures of Kanji characters by dividing into basic elements. Working with the Kanji elements, we found that we can create large number of Kanji characters by compounding small amount of basic elements. This is a very useful method to create many Kanji fonts and to investigate many Kanji characters.

１．e漢字の開発

東洋学文献全体をカバーできるだけの，多数の漢字フォントを作成した．それを，電子漢字という意味で「e漢字」と呼ぶ．

現在，日本最大の漢和辞典である諸橋大漢和辞典は，約５万字の親字を持つ．その字数に，現代中国の略式漢字（簡体字），日本独自の漢字（国字），新字などを加えて，約57000字を当初の開発目標とした．

多数の漢字を一字一字作成すると，そのコストはあまりにも大きくなる．そこで，多数の漢字を効率良く，低コストで作成するために，漢字を部品の組み合わせで生成する手法を採用した．ただし，漢字を機械的に細分するのではなく，字形として意味のある部品単位に分解するよう工夫した．その工夫によって，フォント作成の効率化をはかりつつ，漢字研究に役立つ，漢字の構造的データを得ることができた．

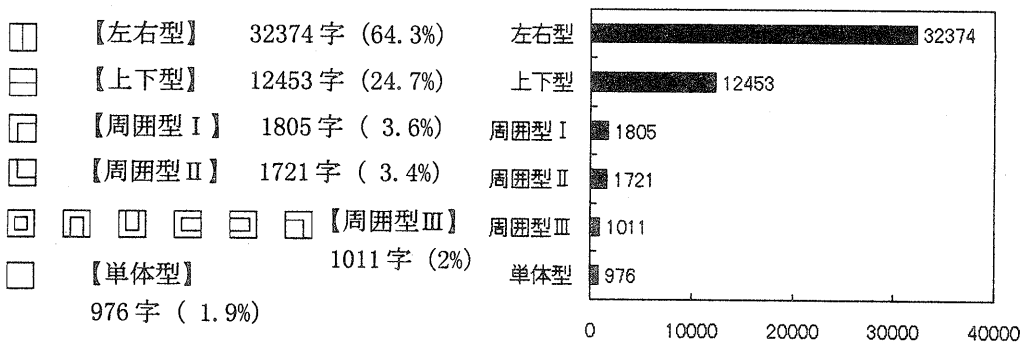
2．漢字の分解

最初に，漢字の構造に着目し，漢字を部品に分解することを考える．

「漢字の分解」という場合，どこまで分解するか，何が最小単位かが問題になる．これまでも漢字を細分化したり，縦線，横線にハネ，ウロコなど数種類の“飾り”を付ければ漢字を作字できる，という考え方はあった．しかし，文字としての漢字を考える場合，あまりに分解すると，漢字の持つ意味や音，分類するための手掛かりが見えなくなる．

そこで，機械的に細分するのではなく，漢字の構造や，その成り立ちに基づく，意味のある部品単位までの分解を行なうことにした．後述するように，漢字を２分割するだけでも作業量が大きく減ることがわかった．

まず漢字を２分割して，分割の形状と，その数を調べる．
諸橋大漢和辞典の50340字[注1]の親字漢字を２分割して，分割形状ごとの個数を数えたところ，下図の結果になった．



このように，約98%の漢字が分割可能文字すなわち合成文字であり，分割形状においては，左右型と上下型だけで約89%を占めている．

分割形状が数種類であることは，漢字を構成する部品のサイズが数種類しかないことを意味する．逆に，数万の漢字を形成するためには，部品の字形は明らかに多数が必要となる．
このことから，漢字を生成するためには，多数の部品を準備する必要があるが，個々の部品の組み合わせ型は，少ないことがわかる．

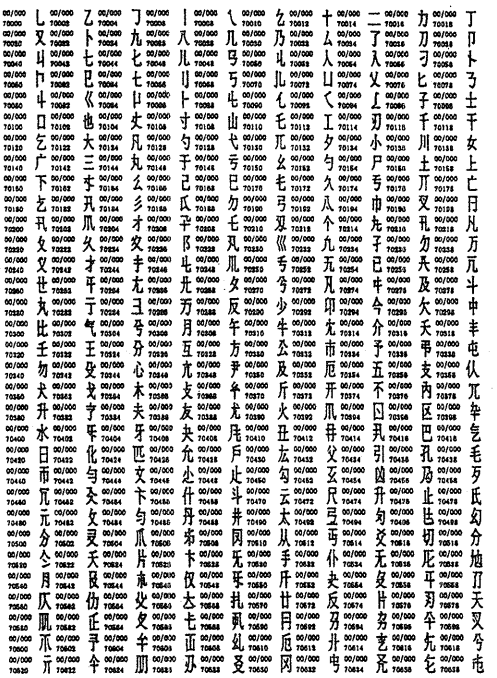
3．部品の変形

部品の総数を得るためには，分割した部品が同じ物かどうかの同定が必要となる．
そのためには，部品の変化を理解する必要がある．

注１）大漢和辞典は，版によって，また重複字，欠字の判断基準によって，親字数が若干増減する．

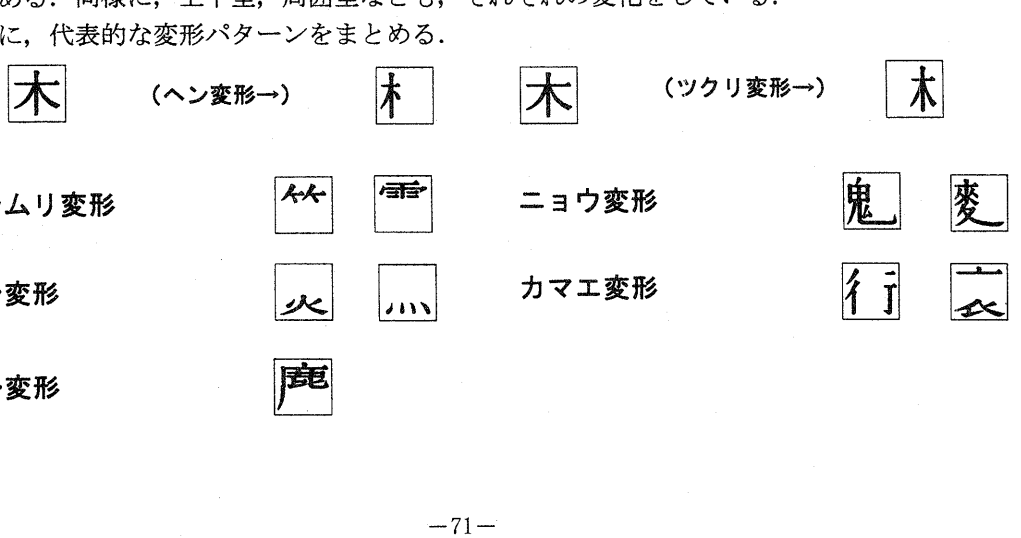
左右分割の，右側（ツクリ）部品をまとめた表の一部を右下図に，左側（ヘン）部品をまとめた表の一部を左下図に示す．





両図を比較すると，同じ形の部品であっても，両者の間には微妙な差異があることがわかる．漢字を構成する部品が，その位置（ヘンやツクリなど）に応じて，本来の形から変化することを，ここでは「変形」と呼ぶ．

ヘン部品の変形，すなわちヘン変形は，右側を短くカットしたり，右上にはね上げるような，特徴的な変化である．一方，ツクリ部品の変形，すなわちツクリ変形は，ヘンのような特異な変化はないが，ヘンと合体するために，本来の大きさが縦細に変化したと考えれば，これも変形である．同様に，上下型，周囲型なども，それぞれの変化をしている．
以下に，代表的な変形パターンをまとめる．



漢字を分割することは，２つの部品の変形パターンを確定し，収集することに他ならない．

変形をどのように規定するかは漢字分解や合成を行う上で，最も重要である．
部品変形を知らなければ，「街」＝「行」＋「圭」，「褒」＝「衣」＋「保」という正しい分割は行えない．変形を規定し，漢字を分割することによって，漢字という合成物の組成を得ることができるのである．

このように，部品は変形した上で他の部品と結合する．その変化は位置（ヘン・ツクリ・ニョウなど）ごとに一定であることから，『部品は，その位置に対して，単一の変形形状しかとらない』という原理が導かれる．これは，漢字を部品から合成するための基本的な認識となる．

４．階層性（基本部品までの分解）

分割された部品の中には，さらに分割できるものがある．その際の分割形状は前記と同様である．２分割した部品をさらに２分割し，その部品をもう一度２分割するというように，部品を細分化していくと，最後には，これ以上は分割できない，分割すると意味をなさない部品にまでたどり着く．それを基本部品と呼ぶ．

まず，分割可能な漢字５６１８９字を２分割した結果，１２１１２個の部品に分かれた．２分割によって，約５分の１に個数が減ったことになる．

さらにくり返し分割した結果，最終的には２４２１個[注1]の基本部品になった．
また，この場合，漢字の階層は，特殊な例外[注2]をのぞいて５階層の分割止まりであることがわかった．

つまり，２４２１個の基本部品をたかだか５回程度合成するだけで，５万字を超える漢字を作り出すことが可能である．

これは，合成による漢字作成が，きわめて有効であることを示している．
漢字という合成文字は，次のような構成になっている．

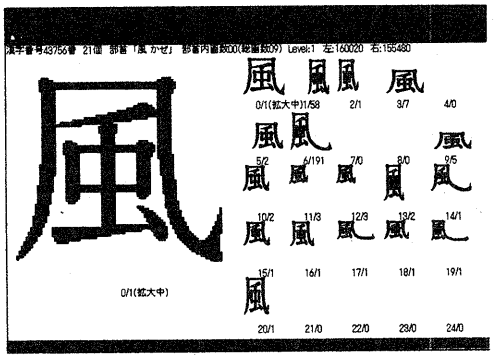
１）漢字は２４２１個の基本部品からなる．
２）基本部品は，変形して他の部品と結合する．
３）結合を，５回程度まで，繰り返す場合がある．
４）結合によって，見かけの部品数は増大し，最終的には，１２１１２個の部品になる．
５）１２１１２個の部品が結合して，５６１８９字の漢字を構成する．

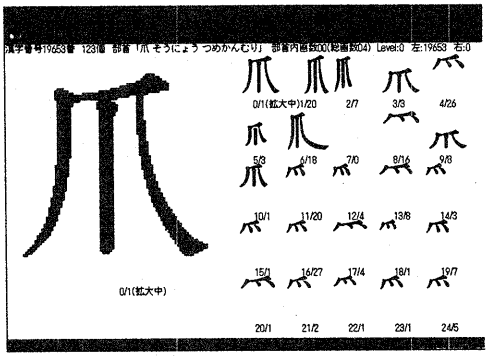
注１）どこまでを基本部品と認識するかで，個数は変わる．
注２）たとえば，“一”を１６個組み合わせたうえに，別の部品も加えた漢字（大漢和番号０１５６４）など．

5．部品の収集・整理

部品には，便宜的に通し番号（「部品番号」と呼ぶ）を付けて，整理する．
また，変形パターンも番号化して，原形＝0，ヘン変形＝1，ツクリ変形＝2，…のようにデータ化していく．ここで，0，1，2，…を変形番号と呼ぶ．

下は，部品「風」「爪」のさまざまな変形パターンを整理している画面例である．





画面例において，大きい部品は，ヘンやツクリとして使われた部品，つまり漢字を２分割して分離された部品である．一方，小さい部品は，部品の中の部品，つまり，何度か分割された結果として分離された部品である．階層が深くなるほど，何度も変形が繰り返されているから，階層に応じて，部品は小さくなる．ここで，２回以上の変形を「重変形」，１回だけの変形を「単変形」と呼ぶ．

合成してから変形しても，個々に変形してから合成しても，同じ結果になることは自明であるから，漢字は，単変形または，重変形部品のイメージ合成である．一つの漢字を構成する，部品のリストを，その漢字の「組成」と呼ぶことができる．

6．部品の統合

右上図の部品「爪」のように，基本的な部品になると，様々な変形パターンがありうるが，なかには，ほとんど差異が感じられない部品もある．これは，重変形において，異なる変形の組み合わせであっても結果的には同じ形になるものや，変形の順番が逆であるにもかかわらず結果的には同じ形になるもの，を示している．

一般に，重変形は，その変形される順番に依存して形が異なるが，現実には必ずしも異なる形状になるとは限らない．その点に留意して，同形状の重変形部品を１個にまとめれば，実際の変形パターンを削減することが可能である．

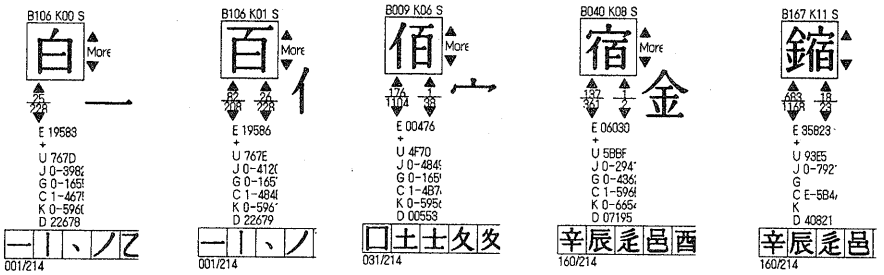
類似の重変形部品を統合することによって，より少ない部品数で，多数の漢字を生成することができる．

7．漢字の合成

漢字の合成は，分解と表裏一体である．漢字の分解作業は，漢字の合成作業用の，変形パターンを作成する作業に他ならない．

前図のように，「爪」部品の変形パターンが，ある程度集まれば，あとは「爪」を含む漢字の「爪」部品は，既に定義された変形パターンのどれかが当てはまる．
したがって，フォント作成作業は基本的な部品パターンを定義する時のみに必要であって，多くの場合は部品の組み合わせだけで，漢字フォントが生成できる．ゆえに，ひと通りの部品パターンが揃ったあとは部品の組み合わせ方，すなわち漢字の組成を得ることに集中するだけでよい．漢字の合成作業は，分解作業の終了とともに，完成しているのである．

下は，合成，検索，確認用に使用したプログラムの画面例である．



ここでは，プログラムが漢字組成式を自動検索して，合成すべき漢字を見つけ出す．
順に，「白」＋「一」→「百」，「百」＋「にんべん」→「佰」，「佰」＋「うかんむり」→「宿」，「宿」＋「金」→「鏹」という，合成過程を示す．

まず，「白」と「一」が与えられると，プログラムは，「白」と「一」を含む漢字組成を検索する．その結果，データベースには「百」しかないことが判明する．
そこで，「百」の組成式は，「百」＝「白」のアシ変換＋「一」のカンムリ変換であることがわかるから，それぞれの部品変換を行って，イメージを合成する．同様に，部品が加えられるたびに，漢字組成にてらして，漢字の候補を表示する．複数の候補が表示される場合は，オペレータが選択して，目的の漢字の組み合わせ方を登録していく．

このように検索と登録を繰り返していけば，複雑な漢字でも容易に作成できる．また，新規に漢字を作り出すことも可能である．

８．組成情報の利用

前図は，漢字合成のために作ったプログラムだが，部品を使った漢字入力方式のデモンストレーションにもなっている．漢字の組成式は，コンピュータへの入力ルーチンに利用できる．

漢字を２分割すると，意味を表わす部品（「意符」と呼ぶ）と，発音を表わす部品（「音符」と呼ぶ）に分かれる場合がある．意符と音符に分けることができる漢字を「形声漢字」と呼ぶ．（漢字の９割以上が形声漢字である．）２分割データを利用すれば，意符または音符並びの漢字リスト作ることができる．

左下は，左右型漢字を右側（ツクリ）部品で，整理した表の一部である．
漢字辞典の部首は，意味的な分類，すなわち，「意符」に基づいている．そして，意符は，左側（ヘン）に位置することが多いために，漢字辞典の並びは「ヘン並び＝意符並び」になっていることが多い．逆に，左上図のような「ツクリ並び」は「音符並び」を意味する．この表は，「同音漢字の一覧表」になっている．

このように，漢字の組成式や部品データは漢字属性のひとつとして漢字入力や漢字検索，さらには，漢字の分類やシソーラス作りに利用できる．

最も基本的なデータ集として，漢字５７４１５字の２分割データ集「漢字典」を，１９９４年に出版した．右下に，漢字典のページ例を示す．

これまで作成した漢字フォントは，京都大学人文科学研究所内のWWWページ「ｅ漢字プロジェクト」http://www.zinbun.kyoto-u.ac.jp/~ekanji/ で，公開している．

ｅ漢字プロジェクトでは，現在，９万字程度の漢字をカバーするために，引き続き，漢字分解データの収集，整理を続行している．さらに，各種の漢字情報のデータベース化を進めている．

