

国立国語研究所学術情報リポジトリ

Classification of Chinese characters by statistical method

メタデータ	言語: jpn 出版者: 公開日: 2017-06-13 キーワード (Ja): キーワード (En): 作成者: 野村, 雅昭, NOMURA, Masaaki メールアドレス: 所属:
URL	https://doi.org/10.15084/00001309

漢字のパターン分類

野 村 雅 昭

0. はじめに

この報告は、「文字・表記の体系的記述のための基礎的研究」の一部として、1978年度に実施した研究にもとづいている。その一端については、すでに、口頭発表(79年5月、国語学会)をおこなった。今回の報告は、それをさらに発展させたものとして記述する予定であったが、種々の事情から、はたせなかった。そこで、今回は、あくまで、研究の第一段階としての実験結果の報告にとどめ、そこからえられた結果にもとづく、第二、第三段階の報告は、後日に期することにした。

1. 目的

表語文字である漢字は、常用されるものだけでも、数千字におよぶ。それをなんらかの観点から、分類・配列することは、文字体系としての漢字の運用上、必要なことである。古来、漢字の分類・配列については、字書の類において、形・音・義の三点からの種々の方法が存在したのも、当然であった。ただし、それらは、検索の便ということを念頭においたものであり、おのずから、配列に主眼がおかれたのは、やむをえないことであった。

問題を現代にかぎるならば、漢字の分類は、教育、印刷、出版、情報処理などの世界で、必要である。それは、複数の漢字の配列の問題にとどまらず、五万字をこえる、いわば無限集合としての漢字を、いくつかの部分集合(セット)としてとらえることを意味する。そして、分類の原理には、にたようなつかわれかたをするものを、ひとまとめにするという条件が、要求されやすい。つまり、漢字の用法や性格による分類である。

このような要求は、伝統的な、形・音・義の三分類でいえば、義(意味)による分類法と、ある面で通じるところがある。たとえば、当用漢字には、「婆」があって「爺」がないとか、「犬」があるのに「猫」がないといった批判は、〈人間〉をあらわす漢字とか、〈動物〉をあらわす漢字とかいった、意味区分を前提としたものであり、用法による分類といっても、さしつかえないだろう。こうした分類を必要とする専門分野もあるにちがいない。

もし、このような意味による分類が、実際の要求と一致するならば、そのこと自体、決して容易ではないが、まだ問題は単純である。しかし、種々の分野からの要求は、多岐にわたっているし、また、漢字そのものに、そうした多様な要求を可能にするような性格がそなわっているため、問題の解決は、決してひとすじなわけではない。

たとえば、「犬」をふくんで「猫」をいれないという方針は、〈動物〉という意味分野に属する漢字をそろえておく(あるいは、いっさいふくまない)といった観点からは、不都合である。しかし、当用漢字の選定方針からいえば、決してそうではない。「猫」には、「ねこ」という字訓と「ビョウ」という字音があるが、字音としてつかわれるのは、「愛猫」「怪猫」ぐらいで、字訓としての用法がほとんどである。それに対して、「犬」の字音である「ケン」は、「洋犬」「野犬」「番犬」「狂犬」「愛犬」などの二字熟語の成分としてさかんにつかわれるほかに、「盲導犬」「狩猟犬」のような三字熟語の成分にもなりうる。つまり、「犬」は「猫」にくらべて、字音としての用法がゆたかであるために、多数の語の表記にもちいられたり、造語性にとんでいたりする。当用漢字の選定方針がこのような点におもきをおいているならば、「犬」があって「猫」がないということは、むしろ当然のことといえる。

しかし、問題は、漢字を分類するのに、どのような観点や尺度があるかということだけではない。表語文字である漢字は、そのことに起因して多様な機能や用法をもっている。そのうちのどれかひとつだけに注目して、分類をほどこすなら、その結果はこうこうであるというだけならば、ことは簡単である。しかし、実際上の要求は種々多様であり、複数の条件が分類選択にあたって付けられるのが普通である。そのためには、漢字のもつ種々の性格や特徴について、

そのどれとどれとの相関がたかいかな、漢字を分類するには、どのような観点や基準によるのがもっとも効率がよいかな、ということを検討する必要がある。

「犬」と「猫」を区別する原理としては、字音としての用法の有無が有効らしいというのは、一種のカンにもとづく推測であって、あるいは、一方が象形文字であり、他方が形声文字であるということが、きいているかもしれないのである。

2. 方法

分類の観点や基準が、二種類あるいは三種類程度であるならば、うえにのべた目的を達成するのは、それほど困難ではない。たとえ、統計的手法をもちいるにしても、それほど努力をはらう必要はないだろう。しかし、それがおおくなればなるほど、ことは容易でなくなる。事実、漢字の分類に要求される条件は、多様であって、たとえば当用漢字のような、漢字のセットは、それらの最大公約数的なものとして存在する。しかし、これまでに、これこれの漢字が、しかじかの条件や基準によってえらばれたという説明がすべての漢字についてなされたことはないし、えらばれてセットにはいつている、どの字とどの字が、そのなかの部分セットを構成するかということも、あきらかにされていない。

その理由のせんさくをすることは、はなしを本筋からはずれた方向にもっていくおそれがあるので、省略する。ただし、その最大の理由がなんであるかということは、推測にかたくない。それは、千字とか二千字とかいった量の漢字を、いくつもの観点から分類し、その相互の関係をあきらかにすることが常人には不可能だということである。漢字を分類する観点は、ざっとかぞえただけでも、すぐに十指をこえてしまう。さらに、こまかくみれば、容易に数十のレベルに達する。それをもととして、千をこえる漢字を分類し、相互の位置づけをおこなうことは、理論的にはかんがえられても、実際には不可能である。かりにできたとしても、それがもっとも効率のよい分類法であるという保証はどこにもないのである。

そこで、統計的な手法のたすけをかりることになる。漢字の分類というよう

な問題は、一見、統計的な手法とは縁どおいものようにかんがえられる。しかし、さいわいなことに、このような、数値情報をもたない、もやもやした対象の分類に適当な手法が存在する。それは、多変量解析の一方法である、林知己夫の数量化理論第Ⅲ類である。この手法は、計算てつづきの複雑さのため、一般への応用ができていたが、コンピューターの普及とともに、各分野でもちいられるようになった。すでに、言語学の世界でも、少数ではあるが、語彙論、方言学、社会言語学などの分野で、応用がこころみられている。筆者も、表記のゆれの分類に、それを用いたことがある。

一般に、多変量解析では、おなじような反応をしめす対象(個体)には、なんらかの意味で共通性があり、また、おなじような性格をもつ対象からえらばれる特性項目は、なんらかの意味で共通性がある、ということが前提となる。また、数量化理論第Ⅲ類は、なまのデータとしては、数値であらわせない、いくつかの属性のカテゴリーをもうけて、それに対する個体(対象)の反応のパターンを、ある方針で数量化して、類似したありかたの個体をそれぞれに数値にかえて分類しようというものである。

いま、問題にしようとする漢字には、種々の属性や機能がある。個々の漢字を個体とみなして、それが属性のカテゴリーにどのように反応するかをしらべ、それを数量化することができれば、類似した性格の漢字をグループとしてとらえることが可能であろう。また、そのような分類をするうえで、どのような観点が有効であるかが判明しよう。

もちろん、それは、漢字という対象が、この手法になじむという前提のもとでいえることである。このようなしわけが、漢字の分類にはふむきできれいな結果がえられないということもありうる。また、えられた結果が、一見、きれいに分類されていたとして、それが統計的には意味があっても、実用のレベルでは、まったく役にたたないものであるかもしれない。つまり、ここでおこなおうとする実験には、漢字をどのように分類するかという目的とともに、漢字のような言語データを、多変量解析の手法であつかうことの可否の検討もふくまれている。

3. 対象と分類の観点

この実験では、国立国語研究所がおこなった、現代新聞の漢字調査（報告56『現代新聞の漢字』）に出現した漢字のなかから、個体となるサンプルを抽出した。サンプルは、かならずしも、調査資料のなかからえらぶ必要はないが、漢字を、純属性的な要因と、計測値にもとづく要因の両面から分類し、比較することを可能にするために、このような抽出法をとった。もし、計測値のみによって分類するならば、かならずしも数量化第Ⅲ類による必要はない。また、数量化第Ⅲ類は、両種の要因が混在していても、さしつかえない。

抽出にあたっては、つぎの各項に留意して、ランダムにおこなった。

- (1) 外的基準としての字種別の個数に、一定のわくをもうけ、それをみtusこと。
- (2) 字種別のそれぞれのわくのなかで、品詞性別の個数の比率に、おおきな差が生じないようにすること。
- (3) ちらばり度のちいさい漢字の特徴的な出現層が、特定の層に集中しないようにすること。

(1)の外的基準としての字種別の個数とは、当用漢字あるいは常用漢字(案)といった、国字施策による範囲をもうけ、それぞれに一定数をわりあててをいう。このような措置をほどこしたのは、漢字の属性による分類が、国字施策とどの程度に一致するものかをみようとしたためである。「当用漢字表」と「常用漢字表(案)」をもとに、5段階の区分をもうけ、各区分ごとに20字をえらぶこととしたが、「当用漢字表」にあって「常用漢字表(案)」にない字は、総数が19字しかないため、10字にとどめ、計90字をえらんだ。90字としたのは、これ以上の個数になると、データの解析にかなりの時間を要すること、および、この程度でも、実験の目的をはたすのに、ほぼ十分な、反応パターンがえられそうだというみとおしをもったことによる。

(3)は、新聞の調査における6種の層別区分のうち、ある特定の層での使用度が高他の層にくらべて、きわだっているばあい、それが特定の層に集中しないように配慮したことをいう。どの層でも、すくなくとも3字は、その層で特徴的につかわれる漢字がふくまれるようにした。

実際の手順としては、(1)の条件によって、下記の五類ごとに、等間隔サンプリングをおこない、必要な字数をえらんだのち、(2)、(3)の条件にてらして、さしかえをおこなった。また、これらの条件をみたすものでも、つぎにのべる、11の観点に対する反応パターンが完全に一致するものは、そのうち1字をのこして、他の字にさしかえた。

○小学校で学習される996字にふくまれる当用漢字表内の字で、常用漢字表案にふくまれる字…20字

○上記の996字以外の当用漢字表内の字で、常用漢字表案にふくまれる字…20字

○上記の996字以外の当用漢字表内の字で、常用漢字表案にふくまれない字…10字

○当用漢字表にふくまれない字で、常用漢字表案にふくまれる字…20字

○当用漢字表にふくまれない字で、常用漢字表案にふくまれない字…20字

11の観点とは、下記のa～kをさす。それぞれに、2～4のカテゴリーをもうけ、個体となる漢字が、そのうちどれか一つの項目にかならず反応するようにした。

a なりたち…いわゆる六書のうち、文字の運用に関する転注と仮借をのぞく4類にわけ、象形と指事をあわせて1項目とした。指事に属する字は、サンプリングのなかには、ふくまれていない。

b 字画数…字画数の複雑さを、総画数によって、3段階に区分した。区分の基準は、当用漢字の総画数の分布をもとにきめた。

c 字体の異同…異体字の有無によって2項とした。

d 造字要素性…その漢字が他の漢字の部分要素になりうるかいないかによって2項とした。たとえば、「工」が「江」や「貢」の部分要素となりうるような例である。ただし、その範囲は、当用漢字の周辺とし、「絹」が「羈」の要素となるようなものまでは、ふくんでいない。

e 音訓…当用漢字音訓表および常用漢字表案にしめされた音訓によって、音のみのもの、訓のみのもの、音訓ともあるものの3項とした。それらの表にない漢字については、表の音訓選定の方針に準じて判定した。たとえば、「唄

一訓のみ」のような例である。

f 品詞性…その漢字をもちいて表記される単語(または、造語要素)の意味的な性格によって、体言類、用言類、相言類、副言類の4項にわけた。副言類に属するものは少数なので体言類のうち、「私」「僕」「誰」の3字を副言類にいった。この3字をぬいたのは、これらの漢字をつかって表記される「わたし」「ぼく」「だれ」などの語が、かなでかかれる傾向がつよいという点で副言類と共通するという理由によるだけで、積極的なものではない。したがって、副言類は、他の三類に対して、いわば「その他の類」ということになる。

以上のa～fは、漢字の純属性的な要因による観点といえる。それに対して以下のg～kは、実際の調査結果にもとづいた観点である。尺度の設定および判定は、現代新聞の漢字調査の数値によっている。

g 語表記率…その漢字をもちいて表記された単語(調査単位)のかず(ことなり数)によって、4段階にわけた。

h 音訓率…その漢字が音または訓のいずれとして使用される傾向がつよいかを、音としての使用率をもとに、4段階にわけた。

i 人地率…その漢字が人名・地名の表記にもちいられるわりあいを、3段階に区分した。

j 使用率…その漢字の調査全体における使用頻度を、4段階に区分した。

k ちらばり度…新聞の調査では、話題による6分野の層別をおこなっているが、その漢字がある特定の層に集中して出現しているか、あるいは、どの層にも平均して出現しているかによって、2項とした。特定の層への集中度の判定は、理論度数と出現度数の比によっておこなった。

サンプルとなった90種の漢字が、以上の11観点に対して、どのような反応をしたかは、表1にしめしてある。観点と基準は、表2にかかげた。

4. 分析

以上のデータを数量化理論第Ⅲ類で解析した結果を、表2および図1にしめす。表2は、第1軸から第3軸までの固有値および各基準ごとのスコアをしめす。図1は、各個体となる漢字の第1軸と第2軸のスコア(表1参照)を平面

座標にプロットしたものである。漢字の右肩にマイナス記号のついているものは、第3軸のスコアが負であることを意味する。(第Ⅲ類による計算は、具体的には、HITAC8250の統計計算ライブラリー(SPL)によった。)

第1軸では、語表記率と使用率の相関がたかいのがめだつ。多数の語の表記にもちいられる率のたかい字ほど、当然のことながら、使用頻度もたかいわけであるが、それが漢字の分類にあたって、もっとも有効な要因となっている点が注目される。図1でいえば、水平方向の右のもののほど使用率がたかく、左のもののほど使用率がひくい。

語表記率・使用率と類似したあたいをもつ基準に、なりたちと造字要素性がある。なりたちのうち象形文字に属するものと、造字要素性のつよいものとは、使用頻度のたかいものとの相関性がたかい。象形文字であることと、造字要素性のつよいこととの関係は容易に理解されるが、これらと使用頻度とのむすびつきは、あまりなさそうにみえる。ただし、一般的な傾向として、象形文字は、それをもととして造字された、会意文字や形声文字に対して、今日でもなお、基本的な文字として使用される可能性がたかいということはいえそうである。小学校で学習される漢字のうち、象形文字は、比較的ひくい学年で履修され、高学年になるにつれて、その比率がちいさくなるというような傾向がみられることも、その一証である。字画数のすくないものが比較的たかいあたいをもっていることも、こうした傾向と関係がありそうである。

第2軸では、音訓と音訓率とがたかい相関をしめし、分離要因として、もっともきいている。普通、第Ⅲ類の分析では、このように相関性がたかいことがはじめから予想される要因を並存させることはのぞましくないが、今回は、実験的なところみなので、あえて、このような方法をとった。ただし、念のために、いずれか一方をのぞいて計算してみたところ、全体の傾向には、ほとんど影響はなかった。

音訓と音訓率は、第1軸でも、純音型・純訓型のものと、音訓ともにつかわれるものとの分離に、ある程度はたらいっている。そして、第2軸では、図1でいえば、上下方向に、訓 $\longleftrightarrow$ 音の対立がみられる。つまり、純音型と純訓型を分離しているわけである。また、音訓交用型は、どちらかという、純音型の

表 1 漢字の数量化 (データとスコア)

漢 字	英 漢 に 対 す る 反 応												ス コ ア			漢 字	英 漢 に 対 す る 反 応												ス コ ア					
	a	b	c	d	e	f	g	h	i	j	k	1	2	3	a		b	c	d	e	f	g	h	i	j	k	1	2	3					
	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3	1 2 3	1 2 3 4	1 2	1 2 3	1 2 3		1 2 3	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3	1 2 3 4	1 2	1 2 3	1 2 3	1 2 3				
田	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.684	-.258	.574	銃	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.633	-.451	-.225		
山	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	1.214	-.187	.177	眼	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.486	-.640	-.060	
子	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	1.402	-.201	.110	岨	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.700	-.664	-.060	
田	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	1.067	.174	.224	且	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.174	.289	-.743	
家	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	1.067	-.433	.024	但	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.057	.582	-.652	
蚕	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.267	-.330	.277	崎	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.365	.548	-.013
王	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.714	-.434	.267	洞	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.356	.774	.437
貞	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.865	.028	-.471	堀	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.324	.857	.153
具	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.354	-.645	-.271	塚	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.355	.557	.263
思	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.613	-.510	-.093	肌	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.003	.890	-.391
亮	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.775	-.314	.226	唇	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.292	-.420	.286
工	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.857	-.204	.547	細	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.034	.461	1.277
湖	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.015	-.545	-.536	靴	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.152	.513	-.077
高	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.243	-.572	-.080	皿	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.658	1.031	-.623
多	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	1.252	-.547	-.127	伍	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.017	.119	.207
若	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	1.330	-.411	-.067	戾	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.536	.517	-.262
旧	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	1.062	-.201	-.044	据	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.532	.350	-.280
私	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.001	-.164	-.257	挟	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.315	-.541	.442
最	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.677	-.125	.026	矯	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.353	-.534	.312
遐	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.734	-.472	-.213	偵	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.584	-.367	-.237
霜	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.034	.267	.877	拐	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.547	-.767	-.470
闊	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.205	.465	.573	頑	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.556	-.672	-.258
蜂	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.567	-.180	-.253	悠	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.457	-.157	-.140
絹	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.132	.562	-.238	朴	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.131	-.176	-.170
琴	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.153	.171	.340	僕	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.536	-.136	-.566
鹽	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.131	.345	.701	藤	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.140	.010	1.077
畔	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.665	-.535	.057	岡	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.067	.671	-.142
芋	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.332	-.074	-.200	捨	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.443	.955	.710
瀧	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.360	.306	-.216	伊	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.216	-.067	-.054
企	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.053	-.238	.511	島	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.116	.478	1.161
込	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.447	.137	.010	頤	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.072	.778	-.607
逕	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.467	.365	-.512	鍋	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.453	.947	.434
調	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.573	-.183	-.223	底	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.644	-.635	-.242
段	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.657	-.531	.365	緩	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.538	-.054	-.104
硬	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.322	.286	.286	唱	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.250	.748	-.072
巨	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.147	.042	-.062	嶺	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.521	.533	-.264
卑	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.758	.201	-.584	貼	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.380	.277	.055
察	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.081	-.545	-.016	輔	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.560	-.307	-.110
又	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.458	-.800	-.536	綴	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.437	-.170	-.003
毀	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.212	.810	-.753	云	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.341	.304	-.510
鳳	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.426	-.471	.052	貴	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.616	.136	-.385
薪	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.367	-.357	.367	碧	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.155	-.358	.371
歌	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	.367	.275	.017	淋	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.505	.045	-.277
	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.258	-.136	.865	妖	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.074	-.077	-.207
	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.463	.077	-.412	遊	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	レ	-.474	.806	-.157

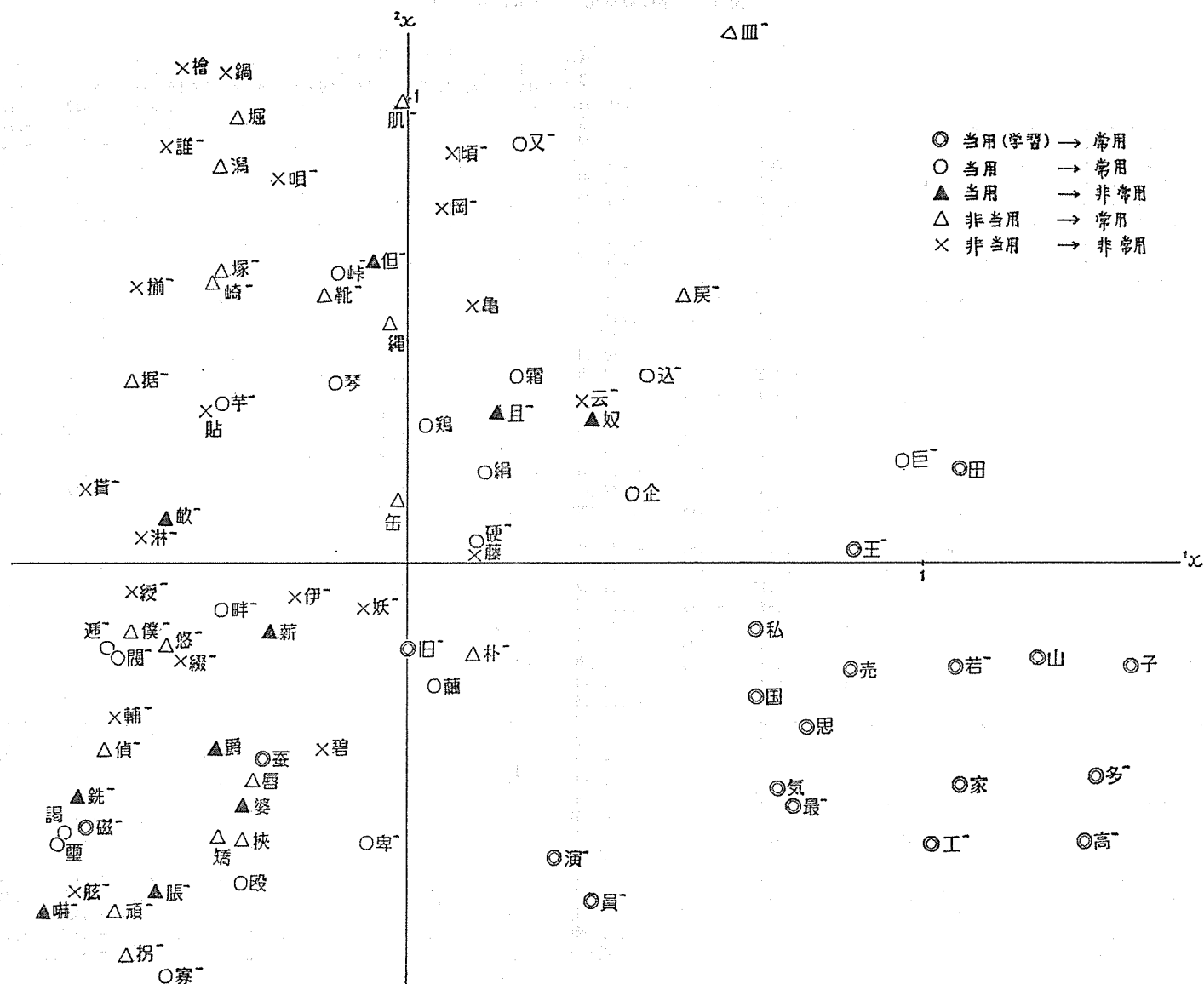


表2 固有値および基準のスコア

観 点	基 準	1x .310 ⁻	2x .231 ⁻	3x .182 ⁻	個体数
a な り た ち	1 象形・指事	2.143 ⁻	.414 ⁻	-.853 ⁻	15
	2 会 意	.727	-.110	-.986 ⁻	14
	3 形 声	-.694 ⁻	-.076	.436	61
b 字 画 数	1 S < 8	1.641	.578 ⁻	-1.276	25
	2 8 ≤ S < 15	-.570	-.509	-.375	48
	3 15 ≤ S	-.803	.589 ⁻	2.936 ⁻	17
c 字体の異同	1 異体字あり	.092	.077	2.525	15
	2 " なし	-.018	-.015	-.505	75
d 造字要素性	1 造字要素になる	2.347	.174	-1.270 ⁻	20
	2 " ならない	-.671 ⁻	-.050	.362	70
e 音 調	1 音のみ	-.664	-1.430 ⁻	-.875 ⁻	32
	2 音調ともあり	1.092 ⁻	-.431	1.620	34
	3 調のみ	-.661	2.517	-1.129 ⁻	24
f 品 詞 性	1 体言類	-.028	.603	.885	46
	2 用言類	-.407	-.817 ⁻	-.423 ⁻	23
	3 相言類	.593	-1.182 ⁻	-.934	14
	4 副言類	.336	1.086 ⁻	-2.560 ⁻	7
g 語表記率	1 30 ≤ W	2.794	-1.312	-.112 ⁻	17
	2 10 ≤ W < 30	1.133	1.942	-1.115	10
	3 4 ≤ W < 10	-.789	.636 ⁻	.909	36
	4 W < 4	-1.127 ⁻	-.741 ⁻	-.729	27
h 音 調 率	1 1.0 = P _h	-1.075	-1.564	-.992 ⁻	24
	2 0.5 < P _h < 1.0	.982 ⁻	-1.136	.597	20
	3 0 < P _h ≤ 0.5	1.046	-.200	2.313	19
	4 P _h = 0	-.508	2.373	-1.189 ⁻	27
i 人 地 率	1 0.5 ≤ P _i	-.269 ⁻	1.353 ⁻	2.277	18
	2 0 < P _i < 0.5	1.731 ⁻	.264	-.196	26
	3 P _i = 0	-.873	-.679 ⁻	-.780	46
j 使 用 率	1 1.000 ≤ P _j	2.729	-1.391	.689 ⁻	16
	2 .100 ≤ P _j < 1.000	.580	1.479	-.094	17
	3 .010 ≤ P _j < .100	-.606 ⁻	1.367	-.310	30
	4 P _j < .010	-1.310 ⁻	-1.626	-.005 ⁻	27
k ち ら ば り 度	1 特徴出現層あり	.068 ⁻	1.120	.101	40
	2 " なし	-.054	-.896	-.081 ⁻	50

注)

$$P_h = \frac{\text{音使用度数}}{\text{音使用度数} + \text{調使用度数}}$$

S = 漢字の総画数

$$P_i = \frac{\text{人名地名使用度数}}{\text{その漢字の総使用度数}}$$

W = その漢字をもちい

て表記された語(短単位)

のことなり数

$$P_j = \frac{\text{その漢字の使用度数}}{\text{全漢字の使用度数の和}} \times 1000$$

ほうにひっぱられる傾向がつよい。

また、使用率は、第2軸でも、非常にたかいものおよび非常にひくいものを、中間のものと区別するのにはたらいっている。使用率の非常にたかいものは、音訓交用型のものがおいから、第2軸における音訓分離の傾向は、使用率のひくいもののほうに顕著である。したがって、図1では、やや右さがりの三角形のなかに、個体が布置していることがみてとれる。

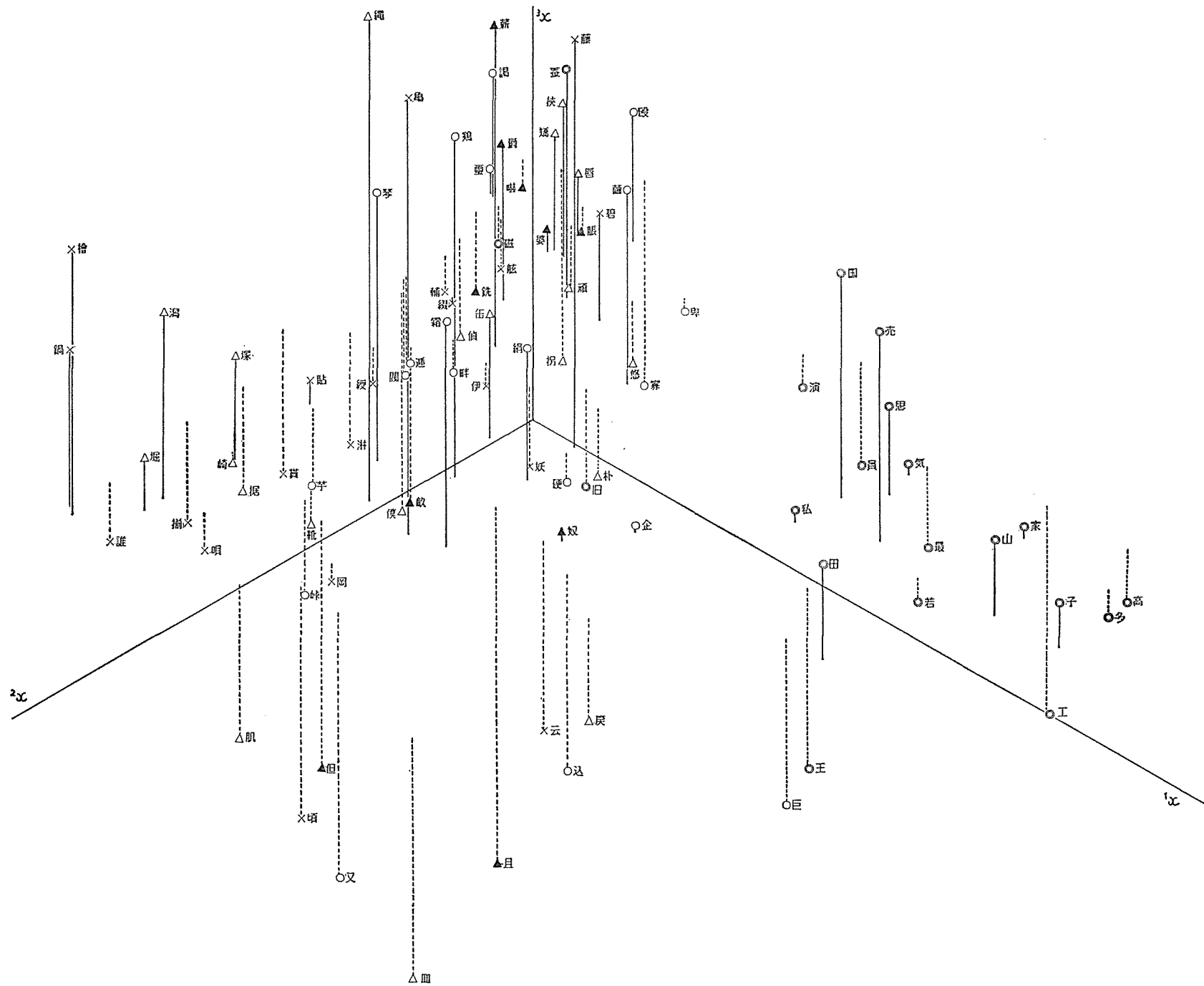
人地率は、第2軸までは、音訓・音訓率とのあいだに、ある程度の対応をしめす。第1軸では、純人地型と非人地型とが、その中間のものと分離される傾向にあり、純訓型と純音型とが、その中間のものと区別されるのと、類似した分布をしめす。そして、第2軸で、純訓型と純音型とが分離するのに対応して純人地型と非人地型が分離する。図1では、純人地型は純訓型とちかいところに布置し、非人地型は純音型と接近している。ただし、第3軸で、純人地型がその他のものと、さらにはっきりと分離し、正のあたいをとるのに対し、純訓型は負のあたいをとり、かならずしも一致してはいない。

第3軸で、人地率のように、分離がはっきりするものには、字画数や字体の異同などがある。第3軸までのスコアによって、三次元空間に漢字を布置したものが図2である。これによって、第2軸までは分離していなかったものがわかれるケースがみられる。たとえば、図1でほぼ同位置の「藤」と「硬」は、上下にへだたっている。これには、人地率や字画数が関係している。また、「奴」と「云」のように、特徴的な要因は指摘しにくい、各要因に対する反応の差が総合されて、この段階で分離するものもある。

もっとも、図2では、第3軸のスコアを上下方向にとっているので、第3軸での差が強調されてみえるが、それを過大視することはひかえなければならぬ。なぜならば、第3軸のスコアは、第1軸・第2軸で分離されなかったものの下位分類としての意味しかもたないからである。おおまかにいえば、全体としては、第2軸までで、分類の目的は、はたされているとみてよい。

これまでにふれなかった観点には、品詞性とちらばり度がある。品詞性は、副言類を他と分離する点では有効であるが、前述のように、その意味はあまりおもくない。全般的には、他の観点ほど顕著な分離要因となっていないといっ

図 2 漢 字 の 布 置 (その 2)



てよい。ちらばり度は、第2軸で、ややたかいスコアをしめすが、全体としては、有効でないといってさしつかえないであろう。

なお、数値およびそれによる図示は省略するが、a～fの純属性的な観点とg～kの計測値にもとづく観点とによって、それぞれ別々に解析をほどこしてみたところ、つぎのようになった。すなわち、純属性的な観点のみによるばあいには、おなじ反応パターンによるかさなりがおおくなるとともに、全体の布置は、上述のものとかなりことなり、分布の特徴を指摘しにくい。それに対して、計測値による観点のみによって解析したばあいには、全体の傾向に、ほとんど差異がみられない。したがって、漢字の分類という点では、計測値による観点を主とすることの有効性は、かなりの程度で保証されるといってよい。ただし、ここでは、その証明にはいたっていない。

5. まとめ

漢字を、その性質や用法によって、いくつかのパターンに分類するという、はじめの目的にたちかえて、今回の実験の結果を整理するうえで、使用率あるいはそれと関係のふかい語表記率が、もっとも有効であるということは、一考にあたいます。なぜならば、使用頻度は、実際に漢字が使用された結果であり、漢字の本来の性格ではないとみなされ、基本漢字のセットをかんがえるうえで、そうした条件を優先させることには、つねに批判がつきまとうからである。

しかし、使用頻度そのものは結果であるとしても、それが分類にあたって有効であるということは、他のもろもろの条件の反映として、使用頻度というものが存在することを意味している。もちろん、厳密にみれば、使用率が有効なのは、その非常にたかいものを他と区別するうえであって、それほどたかくないものでは、音訓率、人地率などの他の条件のほうが効果的であることは、以上の分析からも、あきらかである。したがって、それは、他の条件とのくみあわせで相対的に有効であるという制約のもとで、重視すべき観点である。

この実験の結果から、漢字をパターン分類するならば、つぎのようになる。

- (1) 使用率がたかく、字音・字訓の双方にもちいられるもの

(2) 使用率がひくく、おもに字音としてもちいられるもの

(3) 使用率がひくく、おもに字訓としてもちいられるもの

ただし、パターン分類という点では、この分類は不十分である。なぜならば、この3類のわけかたは、相対的なものであって、実際には、この中間のものが多数存在するからである。そこで、むしろ、このように3類にわけけるよりは、図3のように、これを頂点とする三角形をえがき、これを漢字の分類モデルとして、そのなかに、個々の漢字を位置づけるほうが意味があるとおもわれる。もし、漢字のセットをえらぶ必要があれば、目的に応じて、この三角形を任意にきりとることができよう。

このような結論は、これまで経験的におこなわれてきたものと、あるいは、そうへだたりがないかもしれない。しかし、それでは、ここでおこなったようなてつづきをへて、はじめて、その有効性が証明されるものである。また、こうした方法によって漢字を選択する

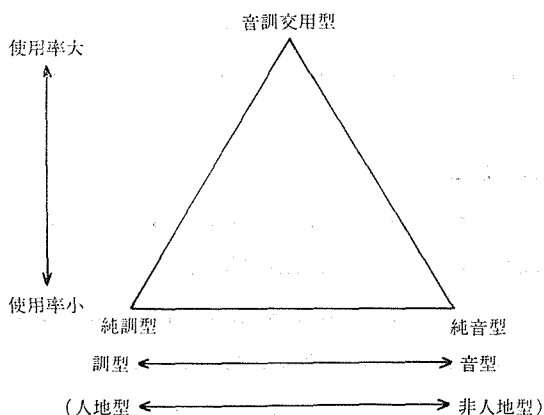


図3 漢字の分類モデル

ことへの批判に対しても、有力な反証となるはずである。そのためには、なお、実験をかさねて、ここにもれている有力な観点はないか、サンプルをいれかえても、その有効性が保証されるか、などの点からの検証が必要である。

さらに、これを国字施策との関係でみるならば、つぎのような指摘ができる。図1・図2でしめした、5種類の字種別の記号の分布をみると、学習漢字(◎印)が他のものと区別されることは、はっきりしている。しかし、その他の字種については、混然としており、この実験の結果とかならずしも一致しているとはいえない。そのことだけを強調するならば、国字施策による字種を選択には、一定の方針がないことになる。

しかし、そうしたみかたが短絡的なことは、いうまでもない。たとえば、図

2で、「込」「戾」「云」の三字は、ほぼおなじところに位置している。このうち、「込」は当用漢字であるとともに、常用漢字案にふくまれている。また、「戾」は、当用漢字ではないが、常用漢字案では採用候補となっている。それならば、「云」も採用候補となってもよさそうであるが、そうではない。これに対するこたえは簡単である。「云」には「言」という同訓異字がある。字音としての用法がせまく、字訓として、他にほとんどおなじ用法をもつ字があれば、その字は採用されない可能性が大きい。この実験で、たまたま、「云」が「込」「戾」とにたあたいをもったのは、「同訓異字の有無」といったような観点のぞかれていることによる。同様のことは、「磁」と「銑」にもいえる。一方が学習漢字であるのに対し、他方は常用漢字からのぞかれようとしている。もし、「磁」を常用漢字にとどめることの正当性を保証するならば、「教育に重要な語彙の表記に使用されるかいなか」というような観点をくわえることが必要であろう。

もちろん、こうした観点をくわえたところで、全体のなかで、それがどれだけ有効であるかは、実験をしてみなければわからない。また、そうしてつづきをへても、なお区別されない漢字、いいかえれば、同一グループとみなしてよい漢字のなかに、ことなる字種のものがふくまれる可能性はある。それに対しては、どれをとり、どれをすてるかという、明確な基準の呈示が要求されよう。そのためには、この実験でこころみたような、観点や条件相互の優先性や、漢字相互の類似性について、科学的なつづきによる分析が必要なのはである。この報告は、そうした目的のための出発点である。

最後に、以上にのべたことを、箇条書きにして、整理してしめす。

- (1) 全体として、数量化理論第Ⅲ類による分類は、基本度のたかい漢字を、その他の漢字と分離するのに有効である。そして、その他の漢字をさらに分類する際の基準をとりだすのにも有効である。
- (2) 基本度のたかい漢字を、その他のものと分離する要因として、使用率・語表記率・なりたち・造字要素性などがはたらいている。
- (3) あまり基本度のたかくない漢字を分類する要因としては、音訓率・人地率がきいている。字画数・字体などの観点も、部分的には有効である。

- (4) なりたち、字形などの純属性的な要因と、使用率・語表記率などの計測値にもとづく要因とでは、後者のほうが全体の分類に効果的である。
- (5) ここでとりあげた観点による分類結果は、基本度のひくい漢字では、国字施策による字種の別とかならずしも一致しない。

(79.10.31)