人文科学とコンピュータ 3－3
（1989. 12. 1）

漢字部首情報からの日本語単語の推定

梅田三千雄

大阪電気通信大学

漢字の部首などの部分構造情報をもとに、日本語単語辞書を用いることによって、どのくらいの単語が決定でき、漢字が推定できるかについて検討した。単語としての文字の性質を定量化するために、ここでは単語確定率と確定文字率の二つの評価尺度を定義した。まず、１１２種類の部首共通文字集合について、部首情報に基づく単語確定率を求め、ランダムに候補文字種を選択した場合と比較した。ついで、「偏」と「つくり」で構成される文字集合の性質を単語確定率と確定文字率により検討した。この結果、漢字の「偏」がわかると約７２％の単語を決定でき、「つくり」がわかると約９５％の単語を決定できることが明らかになった。

Estimation of Japanese Words from the Knowledge about Kanji Radicals

Michio UMEDA

Osaka Electro-Communication University
18-8 Hatsu-cho, Neyagawa-shi, Osaka 572, Japan

This paper discusses how many words can be determined and how many characters can be estimated by utilizing a Japanese word dictionary supposing only radical structures of Kanji characters are correctly recognized. Two evaluation measures such as a word determination rate and a determinable character rate are defined to quantitatively evaluate some characteristics of each character as a word component. First, a word determination rate is calcurated for character sets of common radicals and that is nealy equal to the rate for candidate characters selected at random. Next, two rates are calcurated for the set of characters composed of left- and right-hand radicals. It is shown that about 72% of Kanji characters contained in Japanese words can be determined only from the structure about left-hand radicals and about 95% can be determined from that about right-hand radicals.

１．まえがき

我々が日常の情報伝達媒体として使用している漢字は、表意文字であるが故に文字種が非常に多く、その形も一般に複雑で、しかも形状の極めて類似した文字が多数存在する。また、日本語ワープロがかなり普及した現在でも、漢字は手書きされることが多く、手書きによる文字変形が大きいのも特徴である。これらの理由により、機械による漢字の認識、特に手書き漢字の高精度な認識は、これまでの活発な研究にもかかわらず、パターン認識における難しい課題の一つになっている[１][２]。

これまでに提案されている手書き漢字認識のアプローチには、大別すると、パターン整合法とストローク構造解析法の二つがある。パターン整合法とは、文字パターンそのもの、または文字パターンから抽出した特徴パターンを用いて、標準パターンとの整合処理により、文字を認識するものである。これに対して、ストローク構造解析法とは、文字パターンを構成するストロークセグメント群を抽出し、それらの種類や組み合せ、配置関係などを手がかりに、文字を認識しようとするものである。

手書き漢字認識研究の初期のころには、ストローク構造解析法に基づく文字認識の試みがいくつか報告されている[３][４][５]。つまり、漢字のもつ構造に着目して、抽出したストローク群をもとに、部首に相当する部分構造を理解しながら文字認識を行おうとするものである。しかしながら、これらの研究では、漢字特有の文字種の多さや形状の複雑さの問題を克服することができず、漢字の構造記述や部分パターン抽出の困難さを指摘するに留まっている。

一方、パターン整合法では、文字種の多さは単に辞書規模や処理速度、認識精度などだけの問題であり、認識アルゴリズムの構築にはそれほどの支障にはならない。まして、文字形状の複雑さは、むしろ認識のための特徴として積極的に利用することができる。このため、現在ではパターン整合法に基づく手書き漢字認識が主流となり、多くの研究により、かなり精度の高い認識手法も開発されている。しかし、パターン整合法では、文字全体を一つのパターンとして捉え、全体の形状として最もふさわしいと思われる文字種を認識結果とするため、ある文字の入力に対して、部首などの部分パターンの異なる、思いもつかぬ文字種が認識結果として出力されることがある。

これに対して、我々が漢字を認識し、理解する場合には、漢字のもつ部首などの部分構造の情報が大きな役割を果している。例えば、何か読めない漢字に出くわした時には、部首を手がかりに辞書を検索して、該当する漢字を見つけ出し、その読みや意味を調べる。また、ある漢字を誰かに伝えようとする時には、例えば「ごんべんににん、にんはにんたいのにんです。」とか、あるいは「ごんべんににん、にんはやいばにこころです。」と言うことによって、「認」という漢字を相手に伝達することができる。

さらに、何かある読みづらい漢字に出くわした時には、その漢字の一部の情報をもとに、前後の文字を頼りにして、漢字を推測したり、単語としての意味を理解したりすることができる。また、漢字の一部を度忘れした時など、もっともらしく適当にごまかして書いても、単語あるいは文節、文章として、相手に正しく理解させることができることもある。人間は、このような柔軟な処理をすることによって、現在の文字認識アルゴリズムと比べて、はるかに高度で正確な文字の認識を行っている。

本報告では、文字認識における人間の高度で柔軟な処理のうち、日本語としての単語知識をもとに、漢字の部首情報のみを手がかりとした、単語の推定と漢字の決定について分析する。すなわち、日本語単語辞書を用いることにより、部首などの部分構造の情報しか明らかでない漢字に対して、それを含むどのくらいの単語が推定でき、その文字を決定できるか、つまり単語を構成する文字の単語としての性質を定量化することを試みる。

２．評価尺度の設定

２．１　文字種と単語辞書

単語を構成する文字の単語としての性質を分析するためには、まず対象として扱う文字種を決定し、単語辞書を作成する必要がある。

対象とする文字種は、文字認識アルゴリズムの研究で頻繁に取り上げられる認識対象を参考に、常用漢字１９４５文字とひらがな７５文字（清音字４６、濁音字２０、半濁音字５、小文字４「っ、ゃ、ゅ、ょ」）から成る合計２０２０文字とする。

一方、単語辞書は、学研国語大辞典[６]をもとに、親見出し語に対する正書法として記されているもののうちから、上述の常用漢字とひらがなだけで構成されているものを選択して作成した。ただし、動詞や形容詞などの活用する語は語幹のみとし、ひらがなだけの単語や連語、慣用句、ことわざは除外した。さらに、単語長は２文字以上かつ６文字以下のものに限った。

こうして作成した日本語単語辞書の規模を表１に示す。単語総数は約５２０００語で、その約７割が２文字単語であり、平均の単語長は２．３９文字となる。ただし、上述の文字種のうち、その文字を含む単語が全く存在しないものは「勺、朕、岬、匁、ぢ、ぶ」の６文字であった。

２．２　単語確定率

単語を構成する文字の単語としての性質を議論するため、まず単語確定率[７]なる評価尺度を設定する。

単語確定率（Word determination rate）とは、対象とした単語辞書に含まれる全ての単語を構成する各文字のうち、その文字が正しく認識できなくても、単語を成す前後の文字がわかれば、単語としてその文字と混同する他の文字がなく、単語を一意に確定することができるものの割合で定義する。

この定義によれば、前述の日本語単語辞書における単語確定率は２７．４５％になる。つまり、単語を構成する全ての文字のうち、その約１／４は前後の文字から単語を推定できるという性質をもっていることになる。しかし、これを２文字単語に限って見ると、その単語確定率は０．３１％と極めて小さく、ほとんどの単語を決定することができない。

つぎに、ある共通の性質をもつ文字だけを抽出した文字集合Wにおいて、その集合を成す文字のみに着目した単語確定率を文字集合Wの単語確定率とする。さらに、ある文字「X」に着目して、それを含む単語のうち、その文字がわからなくても、前後の文字から単語を一意に決定できるものの割合を文字「X」の単語確定率と定義する。よって、単語確定率が１００％の文字は、単語を構成する前後の文字がわかれば、単語としての検定を行うことにより、それを含む全ての単語が一意に決まる文字である。また、逆に単語確定率が０％の文字は、このままでは単語辞書を用いる効果が全くない文字である。しかし、何らかの方法によって、その文字に対する候補文字種が限定できれば、単語確定率は大きくなる可能性がある。

そこでいま、各文字に対する候補文字種を全字種のなかからランダムに選択して、対象を限定した場合の単語確定率を求めてみる。すなわち、文字種の総数をＮ個とし、このなかからｎ個の文字がランダムに選択されるものとする。なお、文字「X」に対する候補文字種として、１個は必ず文字「X」が選択され、残りの（ｎ－１）個の文字が（Ｎ－１）個の文字種のなかからランダムに選択されるものとする。

ある単語中の文字「X」について、文字「X」以外にも単語を成す文字の個数をＵとすると、単語を成さない文字の個数はＮ－Ｕ－１（＝Ｖ）となる。したがって、文字「X」以外の文字種が等確率でランダムに（ｎ－１）個選択されると、そのなかに単語を成す文字がｐ個、単語を成さない文字がｑ個含まれている確率Ｐ（ｎ）は、

$$P(n) = \frac{{}_{U}C_{p} \cdot {}_{V}C_{q}}{{}_{N-1}C_{n-1}} \qquad (1)$$

ただし、Ｕ＋Ｖ＝Ｎ－１、ｐ＋ｑ＝ｎ－１

で表される。ここで、

$$p = 0,\ q = n-1 \qquad (2)$$

の時には、単語中の文字「X」は一意に決定されることになる。したがって、その確率Ｑ（ｎ）は、

$$Q(n) = \frac{{}_{U}C_{n-1}}{{}_{N-1}C_{n-1}} \qquad (3)$$

で与えられる。この確率を全字種について求め、各文字種を含む単語数で重み付けすることにより、ランダムに候補文字種を選択した時の、選択文字数ｎに対する単語確定率を導くことができる。

表１　単語辞書の規模

単語長	単語数
2	36398
3	11570
4	3396
5	521
6	77
合計	51962

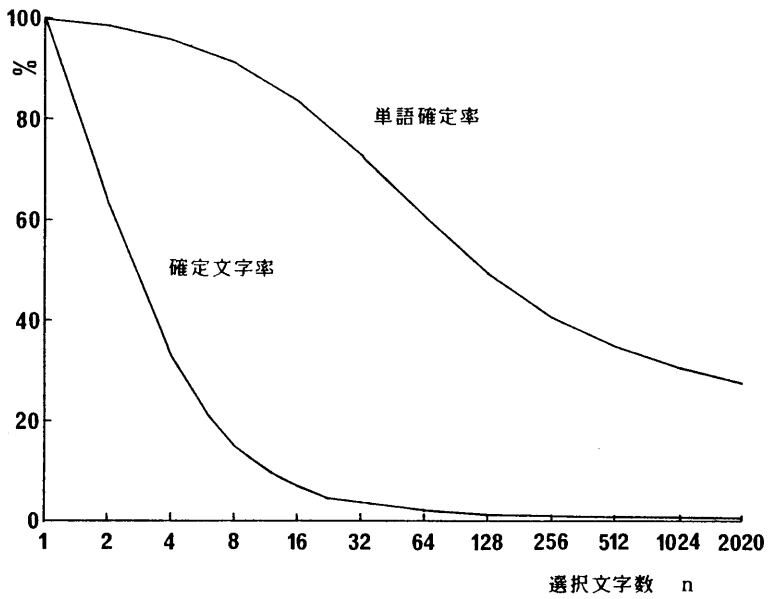

図１　選択文字数に対する単語確定率と確定文字率

前述の単語辞書について、式（３）から算出した、選択文字数ｎに対する単語確定率を図１に示す。これより、何らかの方法によって、各文字に対する候補文字種をランダムに１６個に絞ることができれば単語確定率は約８３％に、さらに８個に絞ることができれば約９１％にまで向上することがわかる。

２．３　確定文字率

単語を構成する文字の性質を議論する、もう一つの評価尺度として、確定文字率[７]を定義する。

確定文字率（Determinable character rate）とは単語確定率が１００％になる文字種の全字種に対する割合、つまりある文字がわからなくても、単語を成す他の文字から、その文字を含む全ての単語を一意に決定することができるという性質をもつ文字の割合とする。さらに、文字集合Ｗにおいて、その集合を成す文字のうち、単語確定率が１００％になる文字の割合を文字集合Ｗの確定文字率と定義する。

前述の単語辞書において、単語確定率が１００％になる文字はわずかに「峠、雰、ぜ、ば、び、ふ、べ、ほ、ゅ、ょ、を」の１１文字しかなく、その確定文字率は０．５４％と極めて小さい。確定文字率も、何らかの方法によって、候補文字種を限定することができれば、大きくなる可能性がある。

ここでも、全字種からランダムにｎ個の候補文字種を選択して、対象を限定した時の確定文字率を求めて

表２　抽出した部首の種類とそれを含む漢字の個数

［偏］				
・汁（１０３）	・化（　８６）	・打（　７８）	言（　５６）	木（　５３）
糸（　５２）	月（　３０）	金（　２８）	・防（　２８）	・忙（　２７）
土（　２６）	・行（　２５）	女（　２４）	・私（　２２）	口（　１９）
石（　１３）	貝（　１２）	日（　１２）	・犯（　１２）	・礼（　１２）
酉（　１１）	日（　１０）	車（　１０）	米（　１０）	・飢（　８）
・初（　８）	・距（　８）	王（　７）	馬（　７）	弓（　６）
山（　６）	舟（　６）	方（　６）	目（　５）	牛（　５）
・次（　５）	・列（　５）			
［つくり］				
・刈（　２７）	・攻（　２５）	頁（　１９）	・役（　１３）	力（　１２）
欠（　１１）	・准（　１１）	・邦（　１１）	寸（　１１）	・印（　９）
斤（　７）	又（　７）	・杉（　７）	月（　７）	寺（　６）
反（　６）	方（　６）	・行（　６）	各（　５）	且（　５）
見（　５）	支（　５）	尺（　５）	主（　５）	召（　５）
青（　５）	東（　５）	皮（　５）	包（　５）	羊（　５）
・悦（　５）	・渇（　５）	・限（　５）	・孔（　５）	・浅（　５）
・場（　５）	犬（　５）	・化（　５）		
［冠］				
・字（　４８）	・芋（　４２）	・笑（　２１）	・介（　１６）	田（　１４）
・文（　１３）	日（　１３）	雨（　１２）	・争（　９）	・究（　８）
・労（　６）	・堂（　５）			
［脚］				
心（　３４）	口（　２３）	貝（　２２）	木（　１８）	・点（　１６）
土（　１６）	日（　１５）	・兄（　１４）	十（　１３）	女（　９）
寸（　９）	月（　９）	田（　９）	・六（　８）	子（　５）
［垂れ］				
・広（　２６）	・尼（　１３）	・病（　１２）	・厄（　９）	・虐（　６）
・戻（　６）				
［構え］				
・囚（　１４）	門（　１２）			
［にょう］				
・込（　５０）	・赴（　５）			

（・印は、その部首がＪＩＳ第１水準漢字にないため、その部首を含む最も簡単な漢字で表したものである）

みる。いま、ある文字「X」に着目し、その文字を含む全ての単語について、単語として文字「X」と混同する「X」以外の異なる文字種の総数をm個とする。選択されるn個の文字のうち、1個は必ず文字「X」が選択され、残りの（n－1）個の文字が（N－1）個の文字種のなかからランダムに選択されるものとする。すると、選択される文字のなかに文字「X」と混同するものがなく、全ての単語が一意に決定できる確率R（n）は、単語確定率の場合と同様に、

$$R(n) = \frac{{}_{N-m-1}C_{n-1}}{{}_{N-1}C_{n-1}} \qquad (4)$$

で表される。したがって、これを対象とした全ての文字種に対して求めることにより、確定文字率を算出することができる。

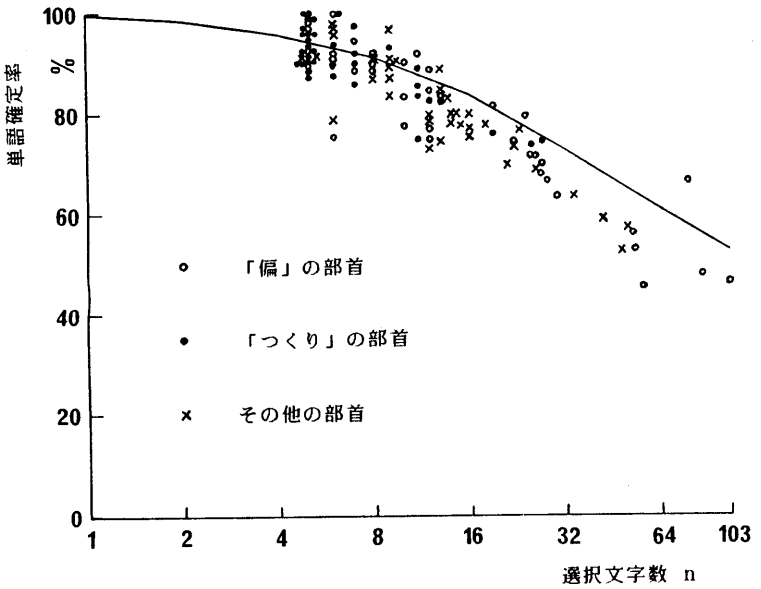

図2　部首共通文字集合の単語確定率

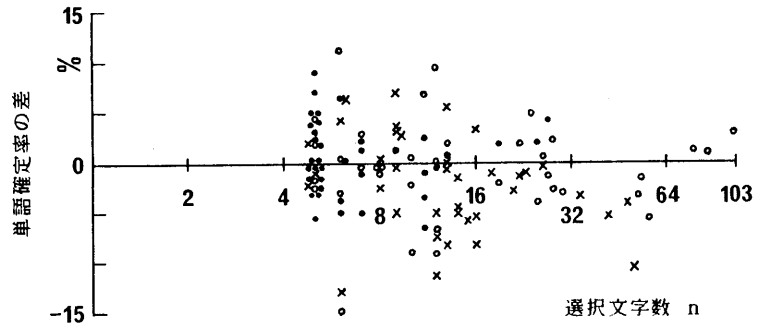

図3　ランダムに候補文字種を選択した時との単語確定率の差

前述の単語辞書について、式（4）から求めた、選択文字数nに対する確定文字率を図1に示す。これより、何らかの方法によって、各文字に対する候補文字種をランダムに8個に絞ることができれば確定文字率は約15%に、さらに4個にまで絞ることができれば約33%になることがわかる。当然のことながら、確定文字率は選択文字数の増加とともに急激に低下していく。

3．部首共通文字集合の性質

3．1　部首共通文字集合の抽出

部首などの漢字の部分構造の情報をもとに、日本語単語辞書を用いることによって、どのくらいの単語が推定でき、文字が決定できるかを検討するために、まず部首を定義し、共通の部首をもつ文字集合を抽出する。

抽出は人手で行い、共通の部首をもつ漢字が5文字以上存在するもののみを対象とした。なお、漢字の特徴として、「偏」や「つくり」、「垂れ」、「構え」、「にょう」に属する部首の定義は比較的明確であるが、「冠」と「脚」に属する部首の概念はあまり明確ではない。つまり、これらの部首では、例えば「くさかんむり」や「たけかんむり」、「れんが」、「したごころ」などのように明確なものもあるが、そうでないものもかなり多い。したがって、ここでは「つくり」に対しては漢字の上から見て、また「脚」に対しては下から見て、同じ部分構造をもつものを部首として定義した。

抽出した部首の種類と、それを部首としてもつ漢字の個数を表2に示す。これより、「偏」に属す部首は37種類、「つくり」に属す部首は38種類、さらに「冠」や「垂れ」などの、その他に属す部首が37種類となる。なお、これらの部首を含む漢字の個数は、延べ1722文字である。

3．2　部首共通文字集合の単語確定率

同じ種類の部首をもつ漢字の集合を部首共通文字集合と呼ぶことにする。表2に揚げた全ての部首

について、部首共通文字集合の単語確定率を求めたものを図２に示す。これは、例えば「さんずいへん」の部首共通文字集合では、「偏」を共有する１０３個の漢字の各々に対して、同じ集合に属す１０３個の漢字が候補文字種として選択された時の、この文字集合の単語確定率を求めたものである。したがって、これは部首などの漢字の部分構造が明らかになった時の、つまり漢字の部首情報からの単語確定率に該当する。なお実線は、全ての文字種について、全字種のなかからランダムに候補文字種を選択した時の単語確定率を式（３）から求めたものである。

これより、部首の種類によってばらつきはあるものの、選択文字数が少ない場合には、両者はかなりよく重なることがわかる。しかし、選択文字数が増加するにしたがって、部首共通文字集合の単語確定率の方が小さくなる傾向にある。これは、選択文字数が比較的多くても単語を決定しやすい単語長の長い単語にひらがなが多く含まれ、それがランダムに選択した時の単語確定率に表れているためと考えられる。

このことをもう少し詳しく検討するために、部首共通文字集合の単語確定率と、それと同じ文字集合の漢字に対してランダムに候補文字種を選択した時の単語確定率との差を求めたものを図３に示す。これより、文字集合によってばらつきはあるものの、全体としては差０を中心として上下にほぼ均等に分布しており、ひらがなを除外して考えると、両者はほぼ一致すると言えよう。つまり、同じ部首の漢字に限っても、単語構成要素としての文字の性質では、他の文字と変わらないことがわかる。

しかし、部首の種類によっては、この差が大きくなるものもある。例えば、「偏」に属す部首では、部首共通文字集合の単語確定率の方が８％以上大きくなるものに、「やまへん」や「かいへん」がある。ちなみに、「やまへん」の文字集合では、それを含む単語数が３２個と少ないものの、部首共通文字集合の単語確定率は１００％になる。一方、部首共通文字集合の単語確定率の方が８％以上小さくなるものには、「ふねへん」や「にちへん」、「こめへん」がある。特に、「ふねへん」の文字集合では、「般」を別にすると、他の「艦、航、船、艇、舶」はよく似た意味をもち、同じような使われ方をするため、単語確定率は１５％も小さくなる。

また、「つくり」に属す部首では、部首共通文字集合の単語確定率の方が８％以上大きくなるものは、わずかに「謁、喝、渇、褐、掲」の漢字に含まれる部首の１個だけで、８％以上小さくなるものはない。さらに、これ以外のその他に属す部首では、８％以上大きくなる部首はなく、逆に、８％以上小さくなるものに「ひとがしら」や「なべぶた」、「あめかんむり」、さらに「営、栄、覚、学、蛍、労」の漢字に含まれる部首の４個がある。特に、後ろの二つは１０％以上も小さくなっている。

このように、漢字の部首などの部分構造に着目して候補文字種を限定しても、部首によってばらつきはあるものの、全体としては、ランダムに候補文字種を選択した時と、単語としての文字の性質にはあまり違いのないことがわかる。つまり、単語としての文字の決定において、漢字の部首情報もそれほど特別の意味をもっているとは言えない。

４．偏・つくり文字集合の性質

漢字の部首のうち、「偏」と「つくり」の概念は比較的明確であり、かつそれらは相補的な関係にある。つまり、「偏」をもつ漢字には、必ず「つくり」に相当する部首が存在する。一方、「冠」と「脚」もまた相補的な関係にあるが、そこに含まれる部首の概念はあまり明確ではない。すなわち、常用漢字の文字集合を対象に、部首を定義して、「冠」と「脚」に相当する部分構造を矛盾なく抽出することは極めて難しい。また、「垂れ」、「構え」、「にょう」には、相補的な部首の考えはなく、部首以外の他の部分構造は漢字によってほとんど全て異なったものになっている。

したがって、ここでは、「偏」と「つくり」を含む漢字だけの文字集合を対象として、部首などの漢字の部分構造の情報から、どのくらいの単語が決定され、その漢字が推定できるかについて検討する。

４．１　偏・つくり文字集合の抽出

常用漢字１９４５文字のうち、「偏」と「つくり」を含む漢字は１０４５文字に及ぶ。ただし、これらの漢字は全て左右に分割でき、かつ「偏」や「つくり」の概念の明確な部首に限定した。したがって、たとえ左右に分割できても、部首の明確でないもの、例えば「八、兆、非、卵、以、川、州、多」などの漢字は除外した。また、「偏」の概念が明確でも、パターンとして左右に分離しにくいと考えられる、「死」も除外した。

こうして抽出した偏・つくり文字集合において、異なる「偏」の種類は１９６個であった。その主なものは表２に示した通りである。したがって、各部首に含まれる漢字の個数は平均５．３個となる。同じ部首をもつ漢字が最も多く存在するものは「さんずいへん」で、１０３個にも及ぶ。ついで、「にんべん」の８６個、「てへん」の７８個の順となる。「偏」は特定のものに集中しており、上位の１０部首だけでも全体の半分を越え、かつ上位の１９個を取れば漢字の個数は７００個に達する。ちなみに、その部首の漢字がただ一つしか存在しないものは１２６個であった。

一方、異なる「つくり」の種類は、表２に掲げた部

首を主なものとして、合計５２０個であった。したがって、各部首に含まれる漢字の個数は平均２．０個になる。最も多くの漢字が存在するのは、「りっとう」の２７個である。ついで、「ぼくずくり」の２５個、「おおがい」の１９個の順となっている。「つくり」は極めて種類が多く、しかもばらついた分布になっている。ちなみに、上位の９８個で漢字の個数がやっと５００個に達し、その合計が７００個になるには上位１９５個の部首が必要である。なお、ただ一つの漢字しか含まない部首は３０５個であった。

４．２　偏・つくり文字集合の単語確定率

偏・つくり文字集合を成す各漢字に対して、それと同じ部首をもつ全ての漢字が候補文字種として選択される場合の、この文字集合の単語確定率を求める。

まず、「偏」に着目して、漢字の「偏」の情報が正しくえられたとすると、例えば「さんずいへん」の漢字に対しては同じ部首をもつ１０３個の漢字が、また「にんべん」の漢字には８６個の同じ部首をもつ漢字が、各々候補文字種として選択されることになる。したがって、「加」のようにただ１個の漢字しかない部首では、他に候補文字種が存在しないため、この漢字を含む単語は全て決定されることになる。このようにして、偏・つくり文字集合に含まれる全ての漢字に対する候補文字種を設定することにより、この文字集合の単語確定率を求めることができる。前述の単語辞書において、「偏」に着目した偏・つくり文字集合の単語確定率は７２．４３％であった。すなわち、漢字の「偏」が何であるかがわかることにより、それらを含む単語の３／４近くが決定できることになる。

同様に、漢字の「つくり」の情報が正しくえられたとすると、この文字集合の単語確定率は９４．５７％になった。つまり、漢字の「つくり」がわかると、それを含む単語の約９５％を決定できることになる。単語辞書の他に、さらに文節や文脈の情報をも利用すると、ほとんど全ての単語を決定できる可能性がある。

４．３　形状類似文字集合の単語確定率

つぎに、文字全体の形に着目して、形状類似文字を候補文字種として選択した時の、単語確定率について検討する。形状類似文字集合を機械的に抽出する確固たる理論はないが、特徴抽出が単純明快で、かつ手書き漢字に対して比較的高い認識率のえられているものが文字の形状をよりよく表しているものと考え、ここでは外郭方向寄与度特徴[８]を用いて、形状類似文字集合を求めた。手書き文字パターンは、電総研で作成されたＪＩＳ第１水準手書き漢字データベースＥＴＬ９[９]を使用し、各文字種について２００サンプルでの平均外郭方向寄与度特徴を求め、L_1 距離を尺度として類似文字集合を抽出した。

いま、偏・つくり文字集合の「偏」に着目すると、例えば「さんずいへん」の１０３個の漢字に対しては各々１０３個の候補文字種を、「にんべん」の８６個の漢字は各々８６個の候補文字種をもつため、この集合の漢字全体に対する平均候補文字数は３８．９３個となる。そこで、偏・つくり文字集合の各漢字に対して各々３９個の形状類似文字を選択し、その単語確定率を求めると６５．１８％であった。よって、「偏」に属す部首に着目した前述の単語確定率の方が、漢字の形状に基づいて同数の類似文字種を選択するよりも、約７．３％大きくなり、より多くの単語を決定できることがわかる。

同様に、「つくり」に着目すると、偏・つくり文字集合の漢字に対する平均候補文字数は４．７０個になる。そこで、各漢字に対して４～５個の形状類似文字種を選択して単語確定率を求め、それから内挿すると、この単語確定率は約９４．０％になった。つまり、「つくり」についても、全体形状に基づいて類似文字種を選択するより、同じ部首の漢字を選択する方がわずかではあるが単語確定率は大きくなることがわかる。

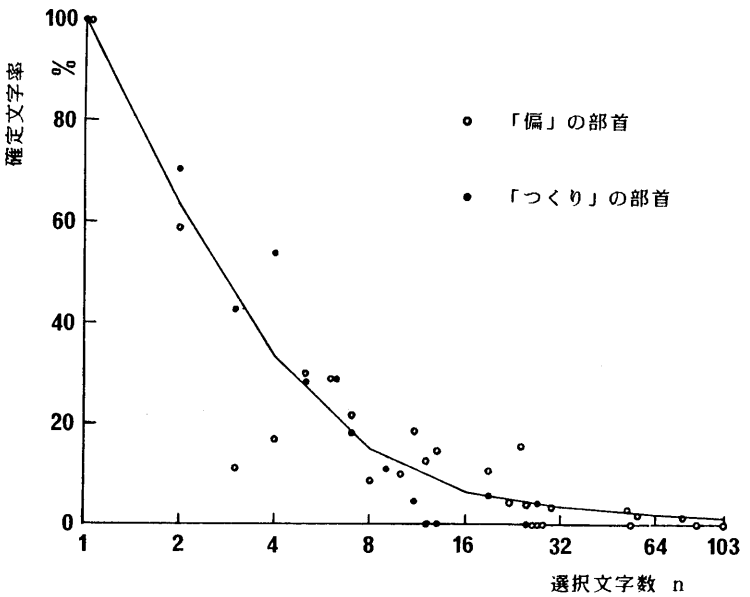

図４　部首共通文字集合の確定文字率

4．4　偏・つくり文字集合の確定文字率

まず、「偏」または「つくり」に属す部首共通文字集合の確定文字率について求めたものを図4に示す。これは、同じ部首の全ての漢字を候補文字種としても単語を完全に決定できるものを求め、「偏」あるいは「つくり」のうちで選択文字数を同じにする漢字群を改めて一つの集合とし、確定文字率を算出したものである。また実線は、全ての文字種を対象として、全字種のなかからランダムに候補文字種を選択した時の確定文字率を式（4）から求めたものである。

部首共通文字集合の確定文字率は、算出の対象となる文字数が限られるため、ばらつきが大きくなるものの、ランダムに候補文字種を選択した時のそれによく重なっていることがわかる。つまり、確定文字率においても、同じ部首の文字集合に、単語としての特別な性質の違いはないことがわかる。

つぎに、偏・つくり文字集合の確定文字率について検討する。単語確定率の場合と同様に、まず「偏」に着目し、各漢字に対して「偏」に属す同じ部首の漢字が候補文字種として選択されるとすると、この文字集合の確定文字率は１９．２３％であった。この結果は漢字の「偏」が何であるかがわかることにより、その漢字を含む全ての単語が推定でき、どの単語からでもその漢字を一意に決定できるものが約2割に達することを示している。なお、この場合の平均選択文字数に相当する３９個の文字を全字種からランダムに選択した時の確定文字率は、式（4）より求めると、わずかに２．８５％である。

つぎに、「つくり」に着目して、偏・つくり文字集合の確定文字率を求めると５８．２８％であった。漢字の「つくり」がわかると、6割近くの漢字がそれを含むどの単語からでも一意に決定できる性質をもっていることがわかる。なお、平均選択文字数に相当する４．７個の候補文字種をランダムに選択した時の確定文字率は約２８．０％である。このように、一定個数の候補文字種をランダムに選択するよりも、「偏」または「つくり」に基づいて候補文字種を選択する方が確定文字率ははるかに大きくなる。

５．むすび

本報告では、部首などの漢字のもつ部分構造がわかると、単語辞書を使用することによって、どのくらいの単語が決定でき、文字が一意に推定できるかを単語確定率と確定文字率を評価尺度として定量化した。その結果、つぎの事項が明らかとなった。

（1）１１２種類の部首を定義して抽出した部首共通文字集合の単語確定率は、同じ個数の候補文字種を全ての文字種のなかからランダムに選択した時のそれとほぼ同じになる。ただし、部首の種類によっては、この差が大きくなるものもある。

（2）「偏」と「つくり」から構成されている漢字の「偏」が何であるかがわかると、それらの漢字を含む単語の約７２．４％が決定できる。同様に、漢字の「つくり」がわかると、その単語確定率は約９４．６％になり、極めて高い値を示す。

（3）これらの単語確定率は、漢字の全体形状に着目して、これと同数の類似文字種を一律に選択するよりも大きくなる。特に、「偏」に属す部首についての単語確定率は約７．３％も大きくなる。

（4）「偏」および「つくり」に属す部首を対象とした、部首共通文字集合の確定文字率も、単語確定率と同様に、同数の候補文字種をランダムに選択した時のそれとほぼ同じになる。

（5）「偏」と「つくり」から成る漢字の「偏」がわかると、その確定文字率は約１９．２％になる。同様に、漢字の「つくり」がわかると、その確定文字率は約５８．３％に達する。すなわち、6割近くの漢字がそれを含むどの単語からでも一意に決定できる性質をもっている。

このように、漢字の「偏」が何であるかが明らかになると、単語辞書を用いることによって、約７２％の単語が決定でき、2割近くの漢字については、それを含む全ての単語が決定できる。一方、「つくり」がわかると、約９５％にも及ぶ単語が決定でき、6割近くの漢字については、それを含む全ての単語を決定することができ、どの単語からでもその漢字を一意に決定できることになる。

参考文献

[1]王,真田,手塚:"手書き文字認識について",信学技報,PRL84-81,pp.35-44(昭60-02).

[2]津雲,浅井:"文字認識技術の最近の動向",信学技報,IE88-5,pp.31-38(昭63-04).

[3]小川,手塚:"漢字の階層表現とその認識",信学論(D),J57-D,12,pp.700-707(昭49-12).

[4]吉田:"ストローク抽出法による手書き漢字認識システム",信学全大,1543(昭49-03).

[5]安居院,中嶋,長橋:"部分パターンの位置関係を利用した手書き漢字の表現法",信学論(D),J60-D,12,pp.1109-1116(昭52-12).

[6]金田一,池田編:"学研国語大辞典",学習研究社.

[7]梅田:"単語辞書を用いた文字認識における文字の確定能力",信学論(D-Ⅱ),J72-D-Ⅱ,1,pp.22-31(平1-01).

[8]荻田,内藤,増田:"外郭方向寄与度特徴による手書き漢字の識別",信学論(D),J66-D,10,pp.1185-1192(昭58-10).

[9]斉藤,山田,山本:"ＪＩＳ第１水準手書漢字データベースＥＴＬ９とその解析",信学論(D),J68-D,4,pp.757-764(昭60-04).