2007－CH－73 (3)
2007／1／27

漢字の諸性質の計量言語学的研究(1)

真田治子(埼玉学園大学)・横山詔一(国立国語研究所)

キーワード：漢字, 計量的言語モデル, 画数, 親密度, 協同的言語学

漢字の画数、読み方の種類の数など計量可能な要素を用いて、漢字の性質の関係性を示すモデルの構築を試みた。今回は、特に漢字の性質や働きを示す9つの要素をとりあげ、画数などの客観的な要素と、個々の漢字に対する親密度(なじみ)などの主観的な要素、あるいは言語内要素と言語外要素、など要素の分類を行い、さらに要素間の相関関係を考慮することで、漢字の計量的な性質に関する全体像を描く試みを行った。この漢字のモデルは、語彙のモデルを提示しているドイツの協同的言語学の理論に着想を得たもので、日本語の漢字には表意性という語彙に近い性質があることから漢字のモデル構築が可能ではないかと考えた。また実証データとしてJIS漢字(6,355字)のデータベースで9つの要素の分布型を検討した。

A study of properties of kanji (Chinese characters) by quantitative linguistic analysis (1)

SANADA Haruko* & YOKOYAMA Shoichi**
*Saitama Gakuen University, **The National Institute for Japanese Language

Key words: *kanji, quantitative language model, number of strokes, familiarity, Synergetic Linguistics*

Our study proposes a model to display the relationship among characteristics of Japanese kanji quantitatively expressed. The present paper focuses on nine properties which can be categorized into objective and subjective ones such as number of strokes and familiarity (psychological perception), and can also be categorized into purely linguistic and not purely linguistic ones. Correlations among the nine properties were investigated to show many quantitative aspects of Japanese kanji as a component of the language. The construction of our kanji model was inspired by the framework of Synergetic Linguistics in Germany which proposed a quantitative language model mainly for the word level. We assume that it is possible to employ a similar approach in our kanji model, because Japanese kanji often represent semantic information as words do in, for example, European languages. To demonstrate our model, the empirical data for the nine properties were obtained from a database of the 6,355 kanji characters of the JIS code, and data distributions of the properties are discussed in our paper.

1. 漢字の表語的性質について(真田)

漢字には、いわゆる表意文字といって、字形(表記)・意味・読みの3つの要素が備わっており、英語のアルファベットのような表記と読みだけの表音文字とは違った性質がみられるといわれている。たとえば漢字の字形の選択に視覚的効果や心理的要因が影響を及ぼしていること(笹原 2006、横山 2006a)や、漢字の語基としての造語力や情報の圧縮力なども漢字の性質の一つである。

一方、漢字圏の中でも日本語の漢字にしかない性質もある。訓読みはもともとその字の意味に対応する和語をあてたもので、たとえば「悲・嬉」に、中国での本来の読みや日本での音読みとは別に「カナシイ・ウレシイ」という語を読みとしてあてる。これは1音節が1字1語に対応する中国での漢字の使い方と異なり、「カナシイ・ウレシイ」という読みの「長さ」をもっているとみなすこともできる。

このように、日本の漢字には字形(表記)・意味・読み(音)、さらには長さをもった語としての読み(訓)を兼ね備え、造語ができる性質がある。つまり漢字は文字ではあるが、きわめて語彙的な性質を同時に有した言語成分であるといえる。たとえば語の長さと語の使用率に負の相関があるのと同様に、画数と漢字の使用率にも負の相関がみられる(宮島 1978)。「短い語ほどよく使われる」というのと同じように「字画の少ない字ほどよく使われる」傾向にあるということである。このような考え方は(表音文字に対して)表"形態素"文字と呼ばれたり文字形態素論などと呼ばれたりしていて、日本語における漢字が欧文のアルファベットとも中国語の漢字とも異なる性質があると認識されている(野村 1987)が、その性質を系統的に整理したものはまだない。

2. 漢字の計量的性質のモデル(真田・横山)

このように日本語の漢字には、記号としての文字と語との両方の性質があり、またその性質のいくつかは画数や読みの長さなど計量的なデータとして示すことができる。本研究ではこの計量的な要素に着目して、漢字がもつ表語的な性質の関係性をモデルにできないかと考えた(図 1)。

このモデルは、

・漢字の諸性質のうち、計量可能な要素を使ったこと
・漢字の諸性質の語彙的な性質に着目して、語彙に準じたこと
・言語内要因と言語外要因を組み合わせたこと

の 3 点に特徴がある。

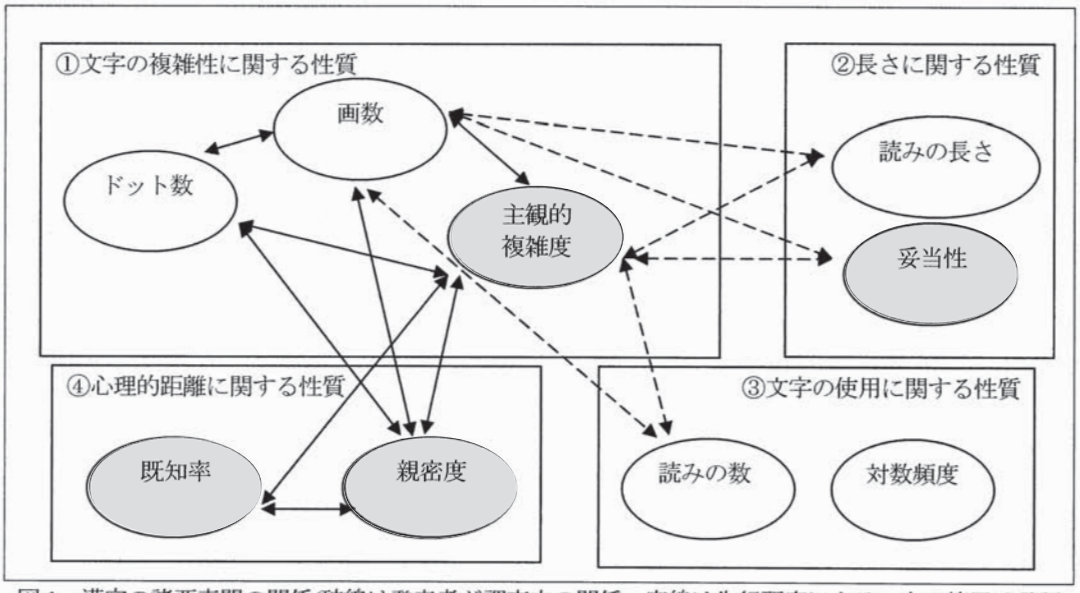

図 1　漢字の諸要素間の関係(破線は発表者が調査中の関係、実線は先行研究による。白い楕円は言語内要因・網掛の要素は言語外要因を示す)

楕円で示したのは画数や頻度など漢字の計量可能な要素で、天野・近藤 1999 の JIS 漢字(6, 355 字)を使った調査を参考にした。このほかにも「辞書における意味の数」(多義性)や語基としての造語力(異なり語漢字カバー率)(田中 1976)などの要素も考えられよう。9 つの要素は、図中では白い楕円の言語内要因(言語自体が持つ性質)、網掛の楕円の言語外要因(話者や社会的要素に関わる要因)に塗り分けて示した。各要素を実際にどのようにデータとして採取するかについては後述する。

さらに9つの要素をその性格に従って、①文字の複雑性に関するもの、②長さに関するもの、③文字の使用に関するもの、④心理的距離に関するもの、の4群に分類した。図中の要素間の関係を示す矢印のうち、破線は本研究で調査中のもので、実線は天野・近藤1999の調査によるものである。

天野・近藤1999の要素間の相関についての調査結果をこのモデルの中に書き込んで、本研究の調査結果とも考え合わせてみると、各群の内の要素間の関係、あるいはある群の要素と他の群の要素との関係には、一定の相関関係があることがわかってきた。たとえば①文字の複雑性の群では、ドット数と画数と主観的複雑度は互いに正の相関があり、画数の多いものは主観的複雑度も多い傾向がある。また①文字の複雑性と②長さ・③使用の群では、複雑群ほど読みの数は少なく読みの長さは長い傾向にある。

実データを用いた要素間の相関関係の分析については次の機会にゆずり、ここではまず、漢字の計量的性質のモデル化の背景と、各要素の基本的性格について考察したい。

3. 協同的言語学 — 言語の計量的モデル(真田)

3-1. 協同的言語学とは

本研究では、漢字のいくつかの要素がもつ語彙的な性質と計量的な性質を生かし、漢字の要素間モデル(図1)を作成したが、この漢字のモデルは、協同的言語学(Synergetic Linguistics)という言語の計量的モデルから着想したものである。

協同的言語学は、ドイツ・Trier大学のKöhler教授が1986年に発表した学説 (Köhler 1986) で、自然科学の協同的理論 (Synergetics) (Haken 1983) にちなんで名付けられた。協同的言語学はHakenの流れをくむものではないが、命名の契機となった自然科学の協同的理論は、粒子のブラウン運動のような個々の無秩序・不安定な現象が協同作用となって、全体として自律形成系をなすことを説明した理論で、たとえば個々の花粉の粒子は別々な動きをしながら全体が大きな一つのブラウン運動を起こすとされる。

本研究で参考にした協同的言語学は、自然科学の協同的理論の考え方を言語になぞらえて構築したもので、言語の中にも一定の自己調節能力(self-regulation、self-organization)があり、花粉の場合と同様に、言語もミクロの視点では一語一語は個別の変化をしているが、マクロの視点では個別の変化は言語全体の大きな動きにつながっていくと考えられている。またこの理論では、「言語変化の過程(Process)」は「調節の要求(Requirement)」が「要素(Variable)」に働きかけるものとして位置づけられ、自己調節を繰り返して言語にとって最適な値が導かれるとされている。言語の自己調節能力は日本語でも観察することができ、たとえば近年新しい外来語が増えているが、では日本語の語彙は無限に膨張し続けているかというと、一方では使われなくなる語もあって日常生活で使われる語彙量はだいたい一定である。

これまでも音韻や語の変化を個別に、経験的・計量的にとらえる研究はあり、Zipfの法則ー「よく使われる語ほど短い」 (Zipf1949)やMenzerathの法則ー「より大きな要素はより小さな部分から成る」(たとえば、長い文はより短い語で構成され、長い段落は短い文で構成される傾向にある)(Menzerath1954) などが知られている。

協同的言語学では、言語変化の現象を巨視的にかつ計量的にとらえるために、言語のさまざまな要素—語の長さ、使用頻度、音素数など—を変数として、種々の言語法則を関数の形で表現し体系化しようとしている。また上述のような計量的な法則性や考え方を見出した先行研究を継承・包含しながら、単に計量的であるだけでなく言語の表層や形態素の変化をその背後にある言語内の「要求」や「調節」といった考え方を用いて、より大きな枠組みで法則化しているところに特徴があろう。そのため、計量的な言語法則を求めながら同時に非常に柔軟性も高いと考えられる。

3-2. 言語変化を説明する成分—要素・要求・過程

協同的言語学は、「調節の要求」が「要素」に働きかけることで「言語変化の過程」が生じるとして、言語変化を説明する事柄をこの3つに区分している。

「要素」には、語のレベルを基本に、語の長さ・使用頻度・言語の語彙量・音素数・類義語の数・多義性などがある。これらは互いに影響関係があり、たとえば日本語の音素数は英語や中国語より少ないため「カガク(科学・化学)」のように同語形の語(同音異義語)が多くなる傾向がある。

「調節の要求」は「生成の最小化」(minimization of production effort: *minP*)や「情報の保持性(冗長性)」(Redundancy: *Red*)、「記号化」(Coding: *Cod*)などがある。「生成の最小化」の要求とは、たとえば話者がことばを発音する労力を最小限にしようとする現象などで、日本語では「ハ」の発音がp音からf音を経て現在のh音になった例がある[1]。

「言語変化の過程(Process)」は「要求」が「要素」にどのように働くかを示したもので、「言語化の最小化」(Minimisation of Encoding Effort: *minE*)や「言語理解の最小化」(Minimisation of Decoding Effort: *minD*)などがある。たとえば(話し手の)「言語化の最小化」という要求は「音素数」という要素を減少方向に導き、結果として「音韻体系の簡素化」という過程が進行するという。逆方向の変化では(聞き手の)「言語理解の最小化」は「音素数」の増加を招き、結果として「音韻体系の複雑化」という過程が進行する。「音韻体系の簡素化」は、日本語では明治時代あたりまで「オクヮシ(お菓子)」と発音されていた語が「オカシ」に変わりカ行音と同一になった例などに見られる。

3-3. 要素・要求・過程の関係とデータへの適用

これらの「過程」は、「要素」と「要求」を使った式の形式で表現することができる。

例① $LS = Cod^{V} * PS^{(-L)}$

「語彙量」(Lexicon Size: *LS*)は「記号化」(Coding: *Cod*)の要求と正の相関(係数 *V*)があり、「多義性」(Polysemy: *PS*)の要素と負の相関(係数-L)がある。つまり言語内で記号化の要求が増大すれば語彙量は同様に増加傾向に動き、多義語が増えれば逆に語彙量は減少傾向に動く。

例② $L = LS^{A} * Red^{Z} * PN^{(-P)} * F^{(-N)}$

「語の長さ」(Word Length: *L*)は「語彙量」(Lexicon Size: *LS*)の要素および「情報の保持性(冗長性)」(Redundancy: *Red*)の要求と正の相関(係数 *A*、*Z*)があり、「音素数」(Phoneme Number: *PN*)および「使用頻度」(Frequency: *F*)とは負の相関(係数-*P*、-*N*)がある。

協同的言語学では、22の要求と15の要素、7つの関係式が提示されている(Köhler 2005)。要求を具体的なデータとしてとらえることは難しいが、要素は実データを適用することができる。その場合、関係式の他の項は一定の係数と考えることもできる。たとえば例②の式から「語の長さ」と「使用頻度」の関係を取り出すには、$LS^{A} * Red^{Z} * PN^{(-P)} = k$ とおくと

例②* $L = k * F^{(-N)}$

と書き換えることができる。この式は「語の長さ」は「使用頻度」と負の相関関係にあることを示しており、上述の「よく使われる語ほど短い」というZipfの法則と同義である。

また要素に実データを適用する場合、様々な解釈が可能である。たとえば「語の長さ」は音韻数、表記した文字数、あるいは日本語なら語を構成する漢字の画数の合計ともとらえることができる。解

[1] ポルトガル人の宣教師が布教のための日本語学習を目的として編纂した『日葡辞書』(1603-1604)には、「Fafa. l, faua」(ハハ. または, ハワ(母))というように、語頭のハ行音は唇音で表記されている(土井・森田・長南 1980)。

釈の自由度の高さは検証の可能性の広がりにつながっており、様々な言語の、様々な言語成分に適用できると考えられる。

3-4. 協同的言語学の漢字の計量的モデルへの応用

協同的言語学の理論は現在も構築中で、言語の性質を計量的なデータに転換し他の言語のデータと比較可能な形にしてしまうことの強みを生かして、ドイツを中心に欧州の言語学者がそれぞれ自国の言語を使ってこの理論を検証している。協同的言語学の理論を直接日本語に適用した研究は、Sanada1999、真田 2002 がある。

本研究ではこの語彙を基本にした協同的言語学の理論を、日本語の漢字に適用してモデル構築を進めたいと考えている。それには、協同言語学が拡張性と柔軟性に富んだ理論である、両者が計量言語学という共通の数学的手法で分析できる、日本語の漢字が語彙的な性質を含んでいて語彙のモデルを適用しやすい、といった背景がある。また漢字圏の漢字でも、このような語彙的な性質をもった表記体系は日本語にしかないことを考えると、言語学的にみても独創的なモデルではないかと考えられる。

4. 漢字の諸要素の分布形(真田・横山)

図 1 に示した 9 つの要素—画数・主観的複雑度・ドット数・読みの長さ・読みの妥当性・読みの数・対数頻度・親密度・既知率—について、天野・近藤 1999 の JIS 漢字(6,355 文字)のデータをもとに分布形をグラフにして確認した。また平均と標準偏差から正規分布を求め、そのカーブも重ねて図示した。

以下、天野・近藤 1999 を『NTT 調査』と称する。主観的複雑度、読みの妥当性、親密度、既知率は『NTT 調査』において 20 代の被験者 24 名に対して行われた心理学的調査の結果である。紙幅の関係から以下に 7 つの要素の分布を示す。

①画数

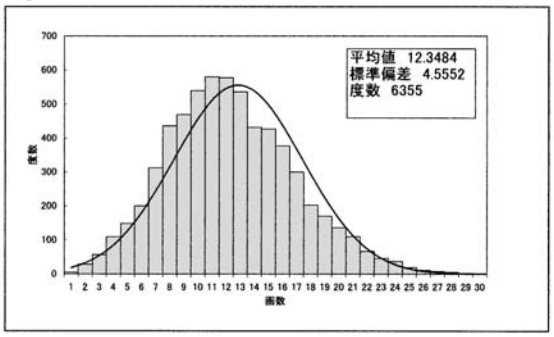

図 2　画数の分布

JIS 漢字(6,355 文字)の画数別分布を図 2 に示す。最小値は 1、最大値は 31 で、左右が比較的対称の山なりになっていることが分かる。一般に日常生活で使用される漢字は 2,000 文字程度といわれており、ここにあげた 6,355 文字の中には珍しい地名・人名等の固有名詞以外にはあまり使われない文字もある。新聞等には 1 年間で 1 回も出現しない文字群を含みながらも，ばらつきなく一定の形状をなした分布になっているが、この点については従来報告されてこなかったと思われる。宮島 1978 では当用漢字(1,850 字、平均 10.3 画)、林 1977 では『新字源』の漢字(9,767 字、平均 12.8 画)の画数別分布を調べている。当用漢字の分布は中央の部分が比較的平坦で 8〜12 画あたりの字数にばらつきが少ない。また『新字源』の分布は右側の画数の多い部分の裾野がのびており、どちらもこの JIS 漢字のような分布とは異なっている。なお、『NTT 調査』には学習漢字・常用漢字・常用漢字外の 3 層に分けた分布図が掲出されている。

②主観的複雑度

『NTT 調査』のうち，実験参加者に「漢字の複雑度を 1(単純)〜7(複雑)の 7 段階で主観的に評定してください」と求めたデータ。事後の内省報では、画数のほかに「隙間がある・知っている・書ける・

意味がわかる・使う」などといった点が判断の基準としてあげられたという。

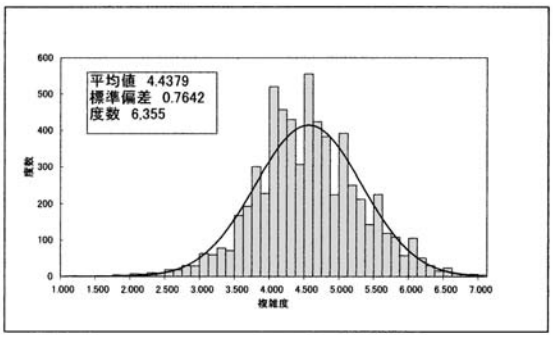

図3　主観的複雑度の分布

文字の複雑度をはかるには、いろいろな方法が考えられる。たとえば、知っている部首や字形が字の部分として組み込まれているとか、あるいは「書」の字のように平行した画が込んでいるなどといったことを点数化することもできる。また先行研究の上述のアンケート調査を見ると、文字の使用との関係も今後考慮すべきではないかと思われる。

③読みの数

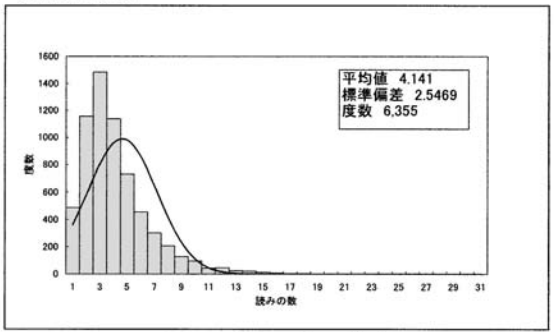

図4　読みの数の分布

『NTT 調査』は、『新明解漢和辞典第 4 版』に記載された読みと『新明解国語辞典第 4 版』から採取した読みの計26,292通りで構成されている。1 文字あたりの平均4.1 通り、最大 31 通りはやや多い印象があるが、これは漢和辞典に一般的にみられる音訓のほかあて字の一部なども含んでいるためと思われる。なお、『NTT 調査』にこの図は掲出されていない。

④読みの長さ

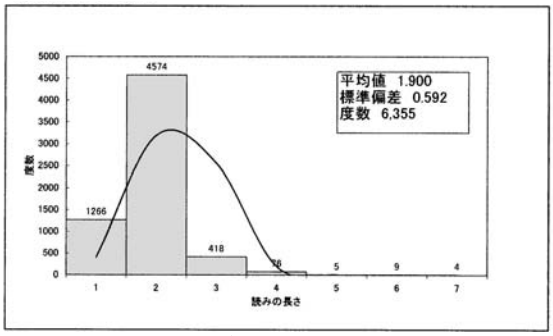

図5　読みの長さの分布

1 節に示した通り、もともと日本語における漢字の読みは中国古来の音の他に、その漢字が意味するものに対応する日本語(訓)であれば、どれでもつけることができる。たとえば「今日」は字音通りコンニチとも読めるし慣用読みでキョウとも読める。また「雨－あめ」のように字形と読みとのつながりがある程度固定的に認識されているものもある。

そこで本研究では、『NTT 調査』をそのまま解析するのではなく、妥当性の指標が最も高い読みをその漢字の読みの代表形とし、その長さの分布を分析した。2 拍が突出しているが、音読みは通常 1 拍と 2 拍に集中すること、このデータベースでは訓読みの送り仮名を長さに含めていないので活用語では語幹の長さが採用される場合が多いこと、などが影響しているかと思われる。今後の研究では音読み・訓読みを分離するか、片方しか読みがない場合はどうするか、あて字の読みを認めるかなどの検討が必要となろう。なお、『NTT 調査』では妥当性が平均値以上の読みについてのみ分布図が掲出されている。

⑤対数頻度

横山・笹原・野崎・ロング 1998 で調査がなされた CD-ROM 版朝日新聞(1993 年)の漢字頻度調査の結果を度数分布として図 6 に示した。一部の高頻度漢字と多くの低頻度漢字から成る典型的な L 字型分布である。JIS 漢字のうち約 3 分の 1 にあたる 2,020 字は度数 0 となっている。横山 2006b によれば 1993 年の使用頻度は 1966 年の新聞の使用頻度と高い相関があり(r=0.95)、この 30 年で大きな変化は見られないという。なお、ここで『NTT 調査』を使用しなかったのは、『NTT 調査』が使用した電子化テキストに文字資料としての問題が発見されていることによる（横山・笹原・ロング・谷本 2001）。

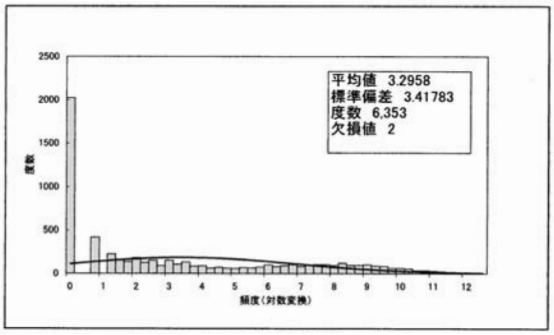

図 6　頻度(対数変換)の分布

⑥親密度

漢字の親密度(なじみ：familiarity)を 1(低)～7(高)の点で評定した心理学的調査。分布は双峰型をなしているが、『NTT 調査』によれば「4.0 以上の山は常用漢字、4.0 以下の山はそれ以外の漢字」で、6.0 以上はほとんど学習漢字であったという。また文字親密度と単語親密度との間に非常に高い相関がみられたという。

上述のように日本語の漢字は単に表記の記号としてだけでなく造語成分として働き、また視覚的な影響も強いので、文字の意味と単語の意味は互いに文字選択や語彙選択に関わっていると考えられる。

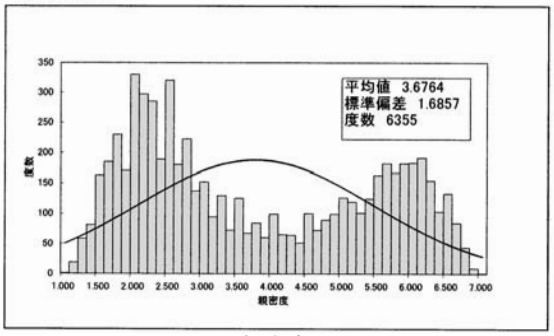

図 7　親密度の分布

⑦既知率

被験者が「漢字として存在することを知っている」ことを問うた調査で、既知の場合は 1、未知の場合は 0 を与える。『NTT 調査』では学習漢字・常用漢字のほとんどが評定 1 に属するとしており、学習漢字・常用漢字とそれ以外の漢字、という双峰型の親密度を裏付ける結果となっている。

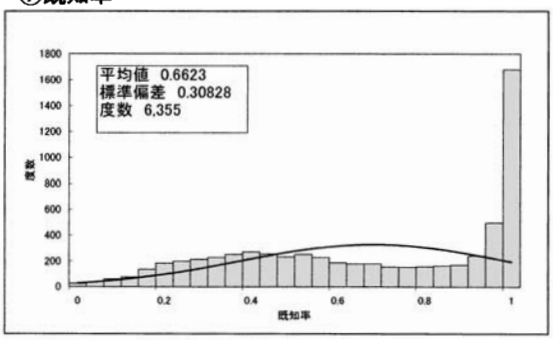

図 8　既知率の分布

5. 今後の展望(真田・横山)

日本語の漢字に特有な語彙的な性質に焦点をあて、計量言語学的な手法によって要素間の関係を明らかにする、漢字のモデルの構築を試みた。今回は漢字の性質のモデル化の意義と、モデル化の模範とした協同的言語学の紹介、さらに先行研究を参考にした 9 つの要素―画数・主観的複雑度・読みの

数・読みの長さ・読みの妥当性・親密度・既知率・対数頻度―の分布形の確認を行った。

さて、読みの妥当性、親密度、既知率などの心理的尺度のうち、親密度については国内外で多くの先行研究が存在し、知見の蓄積が豊富である。英語、仏語、独語など多くの言語で「語」の親密度が対数頻度と高い相関を有するとの報告がなされてきた。この点で、漢字も語と同様の傾向を示す。新聞等で使用頻度の高い漢字は、人間がそれに接触する確率が高くなり、接触確率が高くなると「なじみ」の感覚が生じると考えられている。親密度の高い漢字は、読みやすく、記憶にも残りやすいことが実験心理学や精神物理学の実験で繰り返し実証されている。漢字にまつわる文字生活の実態をモデル化しようとするとき、親密度データは豊富な手がかりを与えてくれるに違いない。なお、親密度は正式な学術用語ではないが、ここでは『NTT 調査』に従った（正式な学術用語は「親近度」）。

漢字のモデルはまだ試作段階であり、そこに組み込む計量可能な要素としていろいろなものが考えられる。たとえば辞書における意味記述の数を多義性の目安としたり、ある漢字が何種類の熟語に使われているか(異なり語漢字カバー率)を造語力の目安としたりすることができよう。漢字の複雑性を計るにも、画数やドット数、主観的複雑度だけでなく、画の長さや交点の数などを数える方法もある。

またこのモデルは漢字を対象としているが、漢字が造語成分であるという点で、語彙のモデルとの接点も考えられる。たとえば、語彙のモデルに属する「語の長さ」と、それを構成する漢字の合計「画数」や「親近度」との相関関係などが考えられる。従ってこの漢字のモデルは、語彙のモデルと二層をなしていると考えることができ、個々のモデル内だけでなく他のレベルのモデルの要素とも関係性を構築できるという点で、拡張性のあるモデルができるのではないかと考えている。

参考文献

天野成昭・近藤公久編著(1999). 『NTT データベースシリーズ日本語の語彙特性』1 巻-6 巻 三省堂

笹原宏之 (2006). 『日本の漢字』岩波書店

真田治子 (2006). 『近代日本語における学術用語の成立と定着』絢文社

田中章夫 (1976). 「漢字調査における統計的尺度の問題」『国立国語研究所報告 59 電子計算機による国語研究Ⅷ』秀英出版, pp.160-191

土井忠生・森田武・長南実 訳 (1980). 『邦訳 日葡辞書』 岩波書店 (『日葡辞書』(1603-1604)長崎学林刊の日本語訳版)

野村雅昭 (1978). 「複合漢字の構造」『文字・表記と語構成』(朝倉日本語新講座 1) 3 章 3 節 朝倉書店, pp.130-144

林大 (1977). 「漢字の問題」『国語国字問題』(岩波講座日本語 3) 岩波書店, pp.101-134

宮島達夫 (1978). 「新字体の画数」『計量国語学』11-7, pp.301-306

横山詔一 (2006a). 「文字の認知単位」『月刊言語』10 月号 大修館書店, pp.36-43

横山詔一 (2006b). 「漢字の使用量」『漢字のはたらき』(朝倉漢字講座 2) 8 章 朝倉書店, pp.169-186

横山詔一・笹原宏之・野崎浩成・エリク＝ロング (1998). 『新聞電子メディアの漢字』三省堂

横山詔一・笹原宏之・エリク＝ロング・谷本玲大 (2001).「新聞漢字調査の現状と将来」『日本語科学』9, pp.33-42

Haken, H. (1983). *Synergetics. An Introduction. Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology.* Third Revised and Enlarged Edition. Berlin: Springer-Verlag.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. (Quantitative Linguistics, vol. 31.)* Bochum: Studienverlag Dr. N. Brockmeyer.

Köhler, R. (2005). Synergetic Linguistics. Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) *Quantitative Linguistics, An International Handbook*. Berlin: Walter de Gruyter, pp.760-774.

Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes.* Bonn: Duemmler.

Sanada, H. (1999). Analysis of Japanese vocabulary by the theory of Synergetic Linguistics. In *Journal of Quantitative Linguistics,* vol. 6-3, pp. 239-251.

Zipf, G. K. (1949) *Human Behavior and the Principle of Least Effort.* Reading, Mass: Addison-Wesley.