漢字の複雑さ

由利高等学校数学班
石川裕真　児玉将大
齋藤陸　　安保裕太

要約

この研究は漢字の複雑さを数値化するために行われた。漢字の要素を調べたり、次元や、相関係数などの考え方を用いて、漢字の複雑さに対する感覚と矛盾しないように式を立てる。これを繰り返してより正確な数値を出すことが目標となる。

１　研究動機と目的

研究のきっかけは班員の小さな疑問で「学年で一番名前の漢字が複雑なのは誰なのか」という疑問から、それを数学を使って解決することはできないだろうかと考えた。

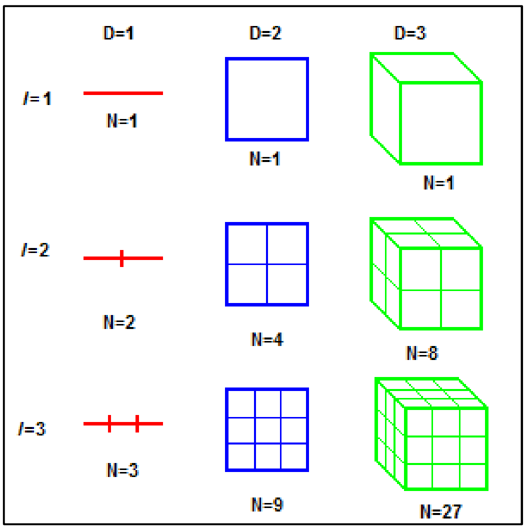
目的は、漢字の複雑さを求める定義式を作りその定義を用いて漢字の複雑さを数値化し比較できるようにすることである。

２　漢字の要素と次元

最初に漢字の複雑さに関わる要素を挙げた。画数、線の長さ、線の密度、同じ部分の繰り返し、対称性などがあげられた。さらに数学の世界では次元が高いほど複雑であるという考え方があり、これを前の要素とあわせて考えていくことにした。

３　次元とは

次元という考え方を取り入れるにあたって次元の考え方について調べた。まず次元には位相次元、相似次元、フラクタル次元というように次元といっても様々あり、それらの違いとは、次元の解釈の広さである。位相次元とは普段私たちが一次元は線、二次元は平面、三次元は立体という考え方のことである。相似次元とは位相次元を大きく解釈したもので、例えばある図形を l 等分したときにできる元の図形と相似な図形の個数を N とする。その数値を $N=l^D$ の定義式に代入し得られた D の値が次元という考え方である。例えば正方形に含まれる線分を二等分するとできる相似な図形は４つでありこれを三等分、四等分したときも



$４=２^2$、$９=３^2$、$１６=４^2$

というように正方形は２次元であることがわかる。そしてこの考え方は位相次元では求められない。図形の次元を表すことにも使うことができる。その例としてコッホ曲線があり、この図形は線分を三等分したときに真ん中を底辺とした正三角形を作り最後に底辺を消し、この作業をできた線分に対して無限回繰り返す。このコッホ曲線は長さが無限であると考えられるので位相次元の考え方ではどの次元にも存在しないことになる。そこで相似次元の考え方を用いて前述した定義式に代入する。

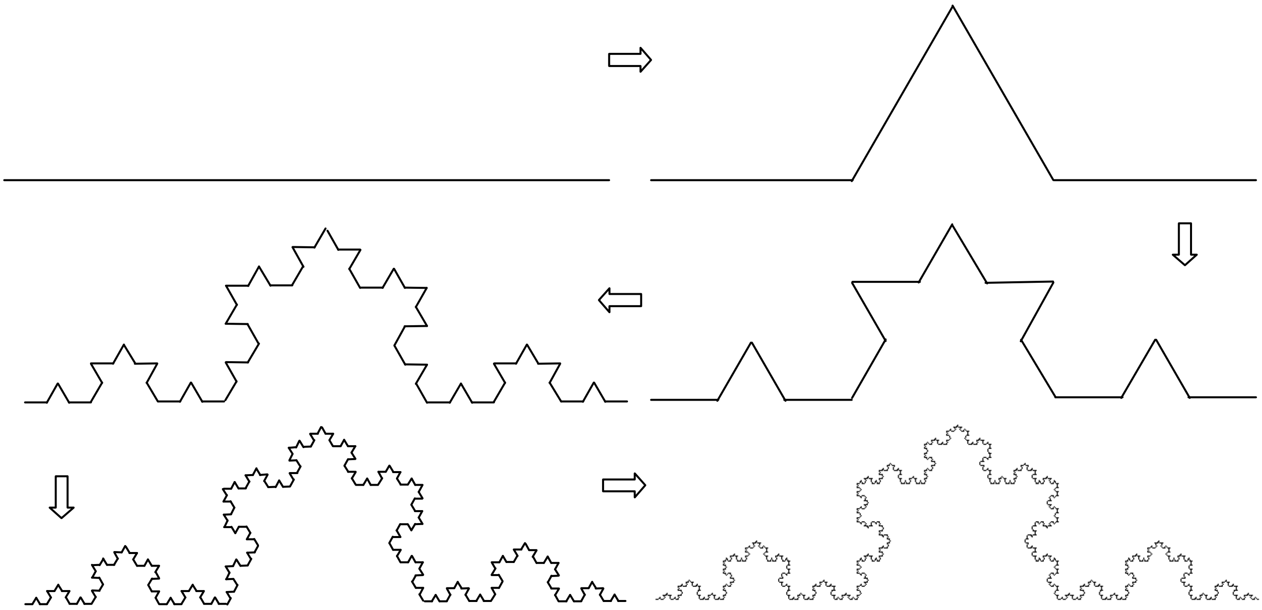
分割数は 3、相似な図形は 4 つということから

$$4 = 3^{D}$$

$$D = \frac{\log 4}{\log 3}$$

$$D \cong 1.26...$$

と求めることができる。下の図がコッホ曲線の５回試行までのものである。



このように複雑な図形の場合、直線で構成されているが、１次元とはならず、非整数の次元となる。線で構成される漢字も様々な角度による交差、様々な変化率の曲線を含むことから、漢字それぞれに特有の次元があるという解釈のもとに、次元を広くとらえた考え方を用いて複雑さを数値化できるはずだと思った。

４　漢字の次元

漢字の次元を求めて複雑さを比較するために、フラクタル次元の求め方の一つであるボックスカウント法という手法を使う。これは次元を求めたい図形をマスの中に納めて、そのマスをさらに小さなマスで分割したときにどれくらいの数のマスに図形が触れているのかを順次調べ、次元を求めるものだ。Ｎを接しているマスの数、εを一マスの大きさとすると

$$N = \mu\varepsilon^{-D}$$

で次元Ｄを定義する。μは誤差調整の役割を果たしているが今回の研究では使わない。この式が成り立っているとき、式変形によって

$$N = \mu\varepsilon^{-D}$$

$$\log N = \log \mu\varepsilon^{-D}$$

$$\log N = \log \mu + \log \varepsilon^{-D}$$

$$\log N = -D \log \varepsilon + \log \mu$$

となり$\log N$と$\log \varepsilon$は傾き－Ｄの直線であらわされる。つまりボックスカウント法を使って数値を算出すると、その数値を通るグラフの傾きから次元を比較できる。またμが関わる値はこの式において切片になるので次元の値には影響しない。次に具体的な漢字の次元を調べた。

５　具体的な漢字で検証

ここでは「引」「創」「龍」「一」の四つの漢字について調べた。この四つにした理由は、複雑さにどのような要素が大きく関係するのか後で調べるために、できるだけ特徴に違いのあるものを選んだ。そして人の感覚では「龍」「創」「引」「一」の順に複雑だと感じるとして検証に移った。

ボックスカウント法を用いて検証した結果が右の表になる。基盤となる正方形の大きさを1024×1024として、左から順に16×16のマスで一辺を64分割、32×32のマスで一辺を32分割、64×64のマスで一辺を16分割、128×128のマスで一辺を8分割したものだ。そして表の数値を対数を利用して縦軸、横軸に表しグラフにした。ここで対数を利用したのはそのままの値をとってグラフを作っても差が大きく表れなかったからだ。次にグラフからそれぞれの漢字の次元（傾き）を最小２乗法を用いて算出した。すると

四つの漢字をカウントした表

引

ε	16	32	64	128
N	865	274	81	33

創

ε	16	32	64	128
N	1258	514	147	44

一

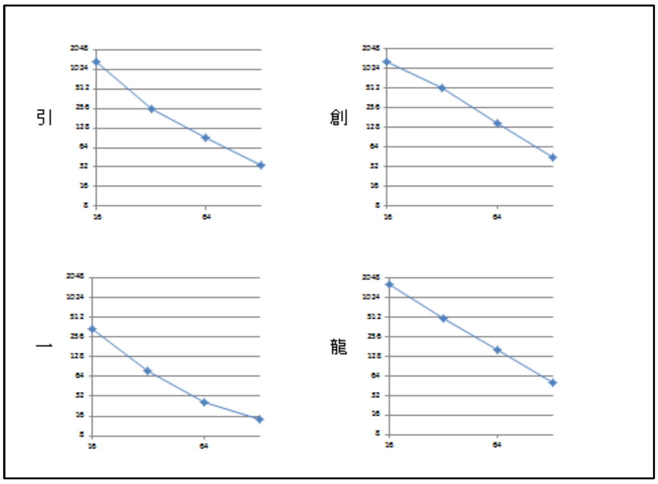
ε	16	32	64	128
N	336	78	26	14

龍

ε	16	32	64	128
N	1648	474	159	50

「龍」…　1.670
「創」…　1.631
「引」…　1.589
「一」…　1.534

となった。しかし、数値にはっきりとした違いが見受けられないので、差を明確にする必要がある。そこで相関係数を用いて数値を調整しようと考えた。相関係数が１もしくは－１に近いほど数値が正確であり信頼できるといえる。それぞれの漢字の相関係数を求めてみると



「龍」…　0.99967
「創」…　0.99748
「引」…　0.99034
「一」…　0.98465

となった。このまま相関係数を使って調整しようとしてもまだ差が出ないので、すべてに共通する小数第一位の「9」を省いたもの、つまり0.9abcdを0.abcdという形に変えて使った。

これを求めた漢字の次元にかけると右のようになった。こうして出た値を「信頼次元」と呼ぶことにした。この時点では人がこの四つの漢字に対して感じる複雑さの順と全く同じといえるので、ほかの要素を含む漢字の信頼次元を新たに調べることにした。

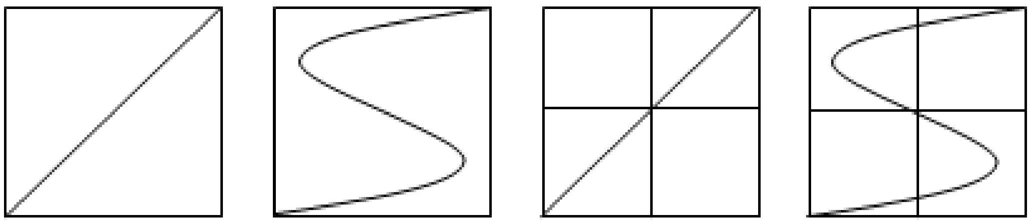
	次元	相関係数	信頼次元
龍	1.670	0.99967	1.664
創	1.631	0.99748	1.590
引	1.589	0.99807	1.435
一	1.534	0.98464	1.298

6　次元に大きく影響する要素

複雑さに大きく関係するだろうと予測した要素は、曲線、同じ部分の繰り返し、対称性、画数の四つだ。新たに調べた漢字は「然」「森」「座」「火」「八」「二」の６つである。

	次元	相関係数	信頼次元
然	1.856	0.9964	1.850
森	1.686	0.9987	1.684
座	1.614	0.9977	1.611
八	1.539	0.9983	1.531
火	1.539	0.9945	1.514
二	1.528	0.9639	1.473

これらの信頼次元は次の図のようになった。調べた結果からわかったことは、曲線が多いと次元が大きくなるということだ。表から読み取ることもできるが、これにはそうなる理由があった。ボックスカウント法を利用した際の直線と曲線を例にあげてみると、粗く数えられているときはどちらも触れているマスの数が同じように扱われていても、分割数が増えたとき、曲線はより多くのマスに触れる場合があるため数値の差が大きくなっている(下図参照)。これが原因で曲線を含む漢字の信頼次元は大きくなるといえる。また、急な曲線と緩い曲線の違いを考えてみると、理由は自明だが急な曲線よりも緩い曲線のほうが直線に近いため次元の増え方が小さい。つまりは曲線が急であればあるほど次元が大きくなるといえることがわかる。繰り返しと対称性については曲線のように顕著に差がみられなかったのでそこまで大きく影響していないように思われた。また、画数もアバウトではあるが、多いほど複雑さが増しているようだった。



7　曲線が与える影響の処理

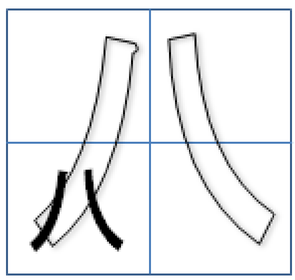
まず曲線について考えてみた。曲線があると人はその図形を複雑に感じるのかといわれるとそう思わないことが多いため、次元を求めたときに曲線が次元を大きくしている分を取り除かなければならない。そこで「八」と「二」の、曲線以外条件がほぼ同じ二つの漢字を使ってそれができないかという仮説を立てた。「八」の曲線は緩い曲線という扱いですすめる。そして次のような式で出した値にたどり着いた。

「八」の信頼次元÷「二」の信頼次元＝1.0393

これを式変形して、実際に漢字の複雑さに反映させるときの形にすると･･･

「八」の信頼次元÷1.0393＝「二」の信頼次元

簡単に説明すると、1.0393 という値で「八」を割ることによって、「八」の曲線が与えた影響を無視した次元、つまり「二」とほぼ同じ次元で扱うことができるということだ。



次に急な曲線について考えてみた。ここでの急な曲線は漢字を四分割したときに、マスを一つしか使わない曲線であり、緩い曲線は漢字を四分割したときにマスを２マス以上使う曲線とする。そして右の図のように漢字を縮小することで緩い曲線になる。急

な曲線は、緩い曲線よりも狭い範囲にあるので、粗くカウントされていると考えることができる。つまり、右図のように急な曲線≅縮小され粗くカウントされた緩い曲線ということになる。これから一番細かい一マス１６ピクセルの数値を抜いた「八」と「二」の信頼次元を使って緩い曲線のときの同様に操作をして急な曲線の数値を出した。それが次の式だ。

１６ピクセルなし「八」の信頼次元÷１６ピクセルなし「二」の信頼次元＝1.0846

今までの作業で得られた数値をそれぞれｃ（緩）、ｄ（急）としてできるだけ正確に反映するために、ある漢字において、緩い曲線の画数を s_1、急な曲線を含む画数を s_2、直線の画数を s_3 として次のような式を立てた。

$$信頼次元 \div \frac{c \cdot s_1 + d \cdot s_2 + s_3}{s_1+s_2+s_3}$$

この式になった理由としては、漢字それ自体に曲線が含まれているかいないかで判断すると正確なものとは言い難いこと。数学的観点からみると画数ではなく曲線の合計の長さで考えるべきだったのかもしれないが、それを調べるには新たな定義が求められることに加えて、私たちが漢字独特の要素をできるだけ反映したいと思ったことなどがあげられる。そしてこの式を使って得られる、曲線が与える影響を無視した信頼次元を準漢字次元とする。この時点ですべての漢字の準漢字次元を調べてみたところ、まだ人の複雑さの感覚に近づいているようには思えなかったため、影響が少ないと思われた繰り返しと対称性についても考えてみることにした。

８　繰り返し、対称性が与える影響の処理

最初に繰り返しと対称性は共通している部分が多いという見解に至った。たとえば「林」のように繰り返しがあることによって、対称性がうまれる漢字が多くある。このことから、この二つの要素は「似通ったパーツによって構成されている」という要素にまとめて考えることにした。また、３分割と２分割の漢字に分けて調べていくことにした。４分割できる漢字は少なく、マイナーのものが多かったので調べなかった。この要素を使った式をたてる際も曲線のときとほぼ同じような考え方で進めた。そして次の式を出した。

３分割･･･「森」の準漢字次元÷「創」の準漢字次元＝1.049

２分割･･･「林」の準漢字次元÷「創」の準漢字次元＝1.021

この値をそれぞれｅ（3 分割）、ｆ（2 分割）として

3 分割できる漢字の準漢字次元÷e　または　2 分割できる漢字の準漢字次元÷f

とすると繰り返しと対称性を無視した準漢字次元となる。これを漢字次元とする。

９　考察

今までに考えた式をまとめると･･･

・相関係数×漢字の次元＝信頼次元

・$信頼次元 \div \dfrac{c \cdot s_1 + d \cdot s_2 + s_3}{s_1+s_2+s_3} = 準漢字次元$

・準漢字次元÷e＝漢字次元

または

準漢字次元÷f＝漢字次元

ここで今までに調べた漢字の漢字次元を求めた。

「然」･･･　1.717
「龍」･･･　1.656
「創」･･･　1.575
「座」･･･　1.553
「森」･･･　1.540
「八」･･･　1.444
「二」･･･　1.443
「火」･･･　1.396
「引」･･･　1.378
「一」･･･　1.272

それが右の表になる。若干の誤差はあるが、ほとんど人が感じる複雑さの順になったのではないだろうか。この表からわかることは次のようなものだった。
・数学的な考え方を用いて漢字の複雑さを表すことはできる。
・式も複雑さを判断する分にはほぼ正確なものを作ることができた。
・画数を細かく考えたり、点を反映する必要があり、そのためには今までの式とは違う視点からのアプローチが必要であると考えられる。

１０　今後の課題

　今後の課題は、ほかの複雑さに影響を与える要素を調べ、それも含めてより正確な漢字特有の次元を求めること。そして学年で一番難しい名前は誰なのかという疑問に取り組むことだ。これからそのような時間が多くとれるかはわからないが、今までの研究で学んだことをこれからの学校生活に生かしていきたい。

１１　参考・引用文献

・フラクタル数学；石村貞夫・石村園子、東京図書、1990
・フラクタル；本田勝也、朝倉図書、2002
・「フラクタル wiki」
https://ja.wikipedia.org/wiki/%E3%83%95%E3%83%A9%E3%82%AF%E3%82%BF%E3%83%AB