ポイヤ数詞表再考

―― 文字統計ならびに漢数詞との比較分析 ――

早田 和弥

The Polya's table of numerals for the ten languages in Europe is analyzed in the context of statistics and information theory. Through a regression calculus of rank-ordered literal data obtained from the table, a statistical law, which would arise from a competition and a frustration effect among letters on the numerals, is found. Subsequently, in an effort to examine a nontrivial proximity between the "length" distribution of the Chinese numerals and those of the European ones, a comparative analysis is made by means of the Hellinger distance. To elucidate the significance of the proximity that has been observed between the Chinese numerals and those of several European languages a statistical hypothesis testing is implemented.

1．まえがき

かつて米国の数学者ポイヤ（Polya, George, 1887-1985)は，表1に示す現代ヨーロッパの代表的な10の言語の数詞を手掛かりにして，2つの言語の一致の度合いが偶然で期待される以上のものか否かを調べる方法（ポイヤの方法）を提出した（安本，1995：21－24)．彼は，これらの数詞の頭文字に注目して2項分布モデルを基に統計的仮説検定を実行し，欧州言語の間に見られる頭文字の一致度は，偶然に依るものではないという結論を導いた(安本，1995：28－29)(1)．本論文では，このポイヤの数詞表を再考し，彼が注目した視点とは全く違う視点からこの表を解析して

表1 印欧系10言語の数詞（ポイヤの表）

	英語	典語	デ語	蘭語	独語	仏語	西語	伊語	ポ語	ハ語
1	one	en	en	een	ein	un	uno	uno	jeden	egy
2	two	tva	to	twee	zwei	deux	dos	due	dwa	ketto
3	three	tre	tre	drie	drei	trois	tres	tre	trzy	harom
4	four	fyra	fire	vier	vier	quatre	cuatro	quattro	cztery	negy
5	five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot
6	six	sex	seks	zes	sechs	six	seis	sei	szesc	hat
7	seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het
8	eight	atta	otte	acht	acht	huit	ocho	otto	osiem	nyolc
9	nine	nio	ni	negen	neum	neuf	nueve	nove	dziewiec	kilenc
10	ten	tio	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz

(安本美典『言語の科学』より） 注：典語＝スエーデン語，デ語＝デンマーク語，西語＝スペイン語，ポ語＝ポーランド語，ハ語＝ハンガリア語

HAYATA Kazuya 札幌学院大学社会情報学部

いる．先ず始めに，表1を構成する100個の数詞に対して，頭文字のみならず尾文字並びにそれらに隣接する文字の出現頻度について回帰分析を行い，単語中の文字サイトと統計則の関係について調べている．次に，表1に示された各言語の数詞の語長分布に着目し，漢数詞のストリング長分布（＝画数分布）との類似度について計量分析を行っている．類似度を判別する為の指標として，ヘリンガー距離を採用している(2)．更に，得られた計算結果に対して仮説検定を行い，この様な類似性は偶然に依るものではないという結論を得ている．

2．文字統計分析

本節では，表1を構成する100個の数詞に対して，文字統計の法則性の有無について調べることにする．表1から得られる各文字（ローマ字アルファベット）の出現頻度を表2に示す．各欄共文字数の合計が100なので，表中の数字がそのまま相対頻度（百分率表示）に対応していることに注意されたい（例：語頭にaが来る相対頻度は3％）．単語中に占める文字の位置（サイト）と統計則の関係を考察する為，ここでは4種類のサイトに着目した．先ず「頭文字」「尾文字」とは，それぞれ語頭，語尾に対応する文字を表す．一方「頭の隣文字」「尾の隣文字」とは，それぞれ頭文字，尾文字の隣のサイトを占める文字を指す．今，フランス語の「3」を意味する〝trois〟を例に採ると，頭文字はt，頭の隣文字はr，尾文字はs，尾の隣文字はiとなる．表2より，いずれのサイトにおいても文字の出現頻度に顕著な偏りが見られるが，この傾向は特に頭文字，尾文字の様な「端文字」よりも「隣文字」について著しいことが読み取られる．例えば，頭文字の最頻値は16（s及びtに相当）であるのに対して，その隣文字ではiに相当する27が最頻値となっている．同様なことが，尾文字とその隣文字についても言える．このことをより定量的に考察する為，表2の各欄から得られる順序統計データに対して回帰分析を行った．結果を表3に示す．表中の数字は当該分布型への適合度｜r｜［式(4)参照］を表す．

分析法の概要は以下の通りである（早田，1997 a）．今，表2第1欄に示した頭文字の場

表2 文字の出現頻度

	頭文字	頭の隣文字	尾文字	尾の隣文字
a	3	2	4	1
b	0	0	0	0
c	5	3	6	2
d	11	0	0	0
e	6	23	18	29
f	7	0	3	0
g	0	1	0	2
h	4	1	0	6
i	0	27	6	11
j	1	1	0	2
k	2	0	0	1
l	0	0	0	1
m	0	0	6	0
n	10	6	13	8
o	6	4	12	3
p	1	0	0	1
q	2	0	1	0
r	0	8	3	9
s	16	1	7	1
t	16	4	8	8
u	3	7	1	7
v	3	1	1	4
w	0	4	0	2
x	0	0	5	0
y	0	3	4	1
z	4	4	2	1
計	100	100	100	100

表3 表2の回帰分析結果
（表中の数字は当該分布への適合度を示す）

	頭文字	頭の隣文字	尾文字	尾の隣文字
EX	0.9922	0.9477	0.9753	0.9684
LG	0.9936	0.9206	0.9919	0.9096
ZP	0.9147	0.9653	0.9233	0.9384
LN	0.9836	0.9576	0.9651	0.9581
N	0.9678	0.8924	0.9633	0.9245
L	0.9265	0.7552	0.9169	0.7097

注：EX＝指数分布，LG＝対数分布，ZP＝ジップ・パレート分布，LN＝離散型対数正規分布，N＝正規分布，L＝直線分布

合を例に採って説明しよう．先ず，当該頻度データを頻度の高い順から並べ換えると，次の様な非負整数列が生成される．

16, 16, 11, 10, 7, 6, 6, 5, ……,
1, 1, 0, 0, 0, 0, 0, 0, 0 (1)

この数列の順序を説明変数xに，各項の値を目的変数zに割り当てると

(x, z) = (1, 16), (1, 16), (3, 11), ……,
(18, 0), (18, 0) (2)

を得る[3]．ここでは，頻度データに内在する統計則の有無を探求することを目的として，変数xとzの変数変換によって新たな変数u，vを定義する[4]．

$$u=[\alpha x+(1-\alpha)\log x]^{p} \quad (3\,a)$$

$$v=[\beta z+(1-\beta)\log z]^{q} \quad (3\,b)$$

ここにα，βは線形関数と非線形関数を統一的に記述する為に便宜的に導入されたパラメータであり，いずれも0又は1の2種類の値のみとるものとする．p，qは対数変換の非線形性の度合いを表す正パラメータである．対数の底は10（常用対数）とする．回帰曲線の関数型として，本論文では次の6種類のものを考える．

EX型）指数分布：
$(\alpha, \beta)=(1, 0)$, $(p, q)=(1, 1)$
LG型）対数分布：
$(\alpha, \beta)=(0, 1)$, $(p, q)=(1, 1)$
ZP型）ジップ・パレート分布：
$(\alpha, \beta)=(0, 0)$, $(p, q)=(1, 1)$
LN型）離散型対数正規分布：
$(\alpha, \beta)=(0, 0)$, $(p, q)=(2, 1)$
N型）正規分布：
$(\alpha, \beta)=(1, 0)$, $(p, q)=(2, 1)$
L型）直線分布：
$(\alpha, \beta)=(1, 1)$, $(p, q)=(1, 1)$

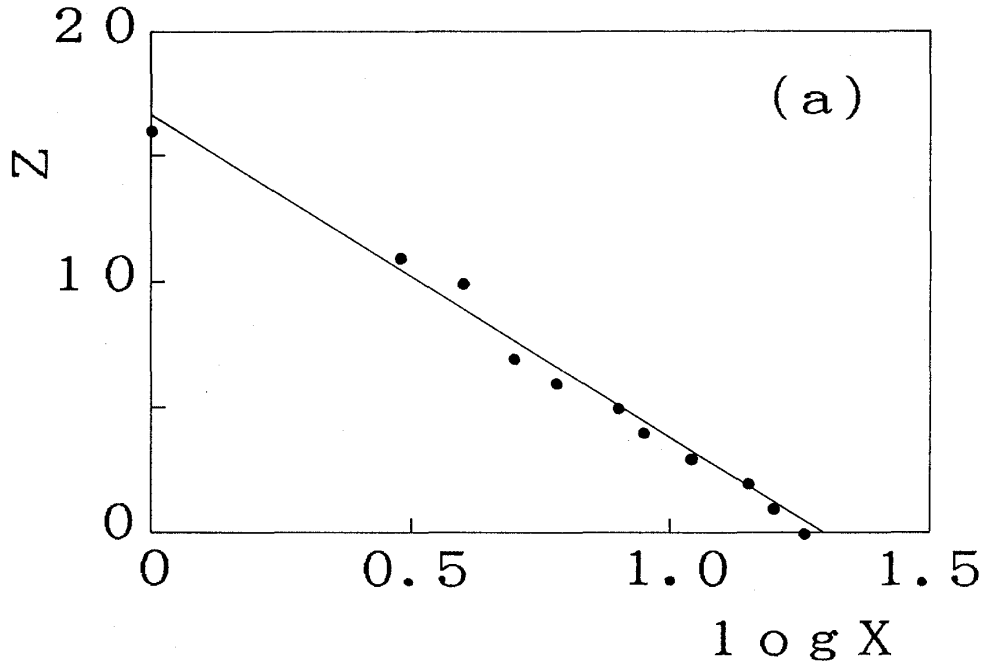
ここに直線分布とは，規範的な線形単回帰モデルで仮定する分布型に他ならない（∵u=x, v=y）．即ち，もし直線分布への適合度[式(4)参照]が他のどの分布の場合よりも高く，更に残差分析に代表される回帰診断にパスしたとすると，当該文字統計は線形単回帰モデルに従うものと判断してよい．

式(2)，(3)より得られる26個の点(u_i, v_i)をuv平面上にプロットし，記述統計学の常套手段（岡本他，1977：21－23）に従い回帰分析を行う．変数変換を施されたデータ(u_i, v_i)の当該分布への適合度は係数

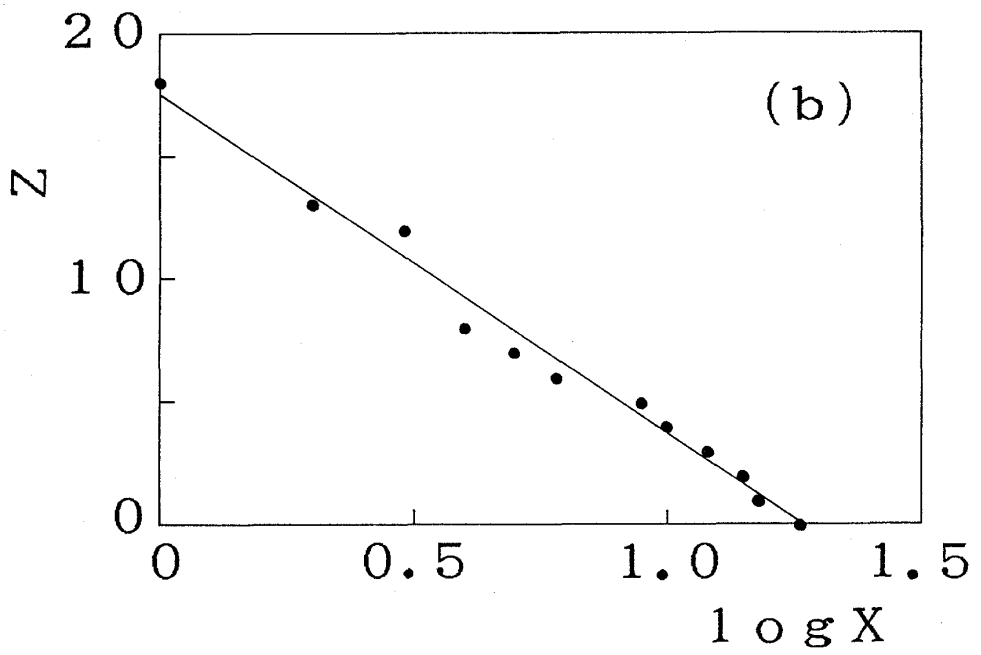
$$r=s_{uv}/(s_u s_v) \quad (4)$$

の絶対値$|r|$によって計量化される．ここにs_u, s_vは各データ変数の標準偏差を，s_{uv}は共分散を表す．尚，rは$|r|\leq 1$という様に正規化されている．本手法の特徴は，言語体系における文字出現頻度に見られる「競合」や「共存」，更にはこれらの「中庸」状態と考えられる「競合的共存」といった，いわゆる「文字生態学」的概念によって当該文字統計を解釈することができる点にある．

表3に示す回帰分析結果から，印欧語族系数詞の文字統計に関して興味深い結論を引き出すことができる．即ち，単語の端のサイトに位置する文字（頭文字と尾文字）については，LG型の分布に最も適合していると言えるが，これらに隣接したサイトに位置する文字（頭の隣文字と尾の隣文字）については最適な分布型を明確に同定することができない．頭文字データと尾文字データのLG型への回帰プロットをそれぞれ図1(a)，(b)示す．これらの平面において，点（黒丸）の見かけの総数は明らかに26より少ないが，これは重複している点（縮重点）が幾つか存在することに依るものである．例えば，図1(a)において，(x, z)=(1, 16)，(6, 6)，(14, 2)，(16, 1)



(a) 頭文字：$|r|=0.9936$；$\hat{z}=16.58-12.94 \log x$



(b) 尾文字：$|r|=0.9919$；$\hat{z}=17.49-13.82 \log x$

図1 対数分布（LG型）への回帰図．目的変数z，説明変数xはそれぞれ文字の出現頻度（表2）とその順位を表す．

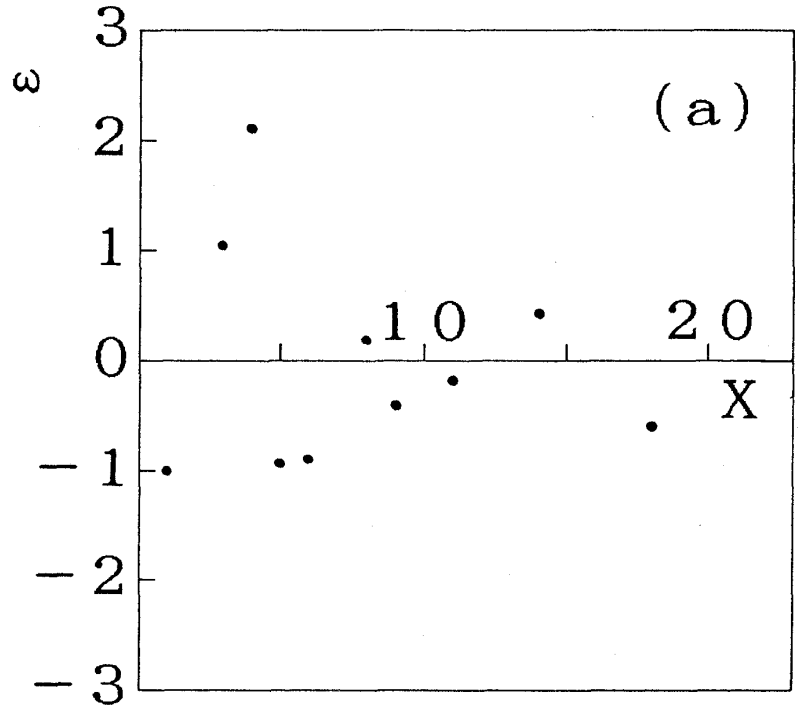
なる点は2重に縮重している．又$(x, z)=(11, 3)$，$(18, 0)$なる点はそれぞれ3重，9重に縮重している．

さて，この様に端文字統計に対してはLG型の分布への適合性が認められたが，はたして当該文字統計はLG則に従うと言えるのであろうか？ ここにLG則とは「LG型の分布に対して適合度が最大で，回帰診断の結果，このモデルに従うと判断される場合に同定される統計則」として定義する．ここでは回帰診断法として，残差分析を採用した．LG分布に対する標準化残差と順位の関係を図2に示

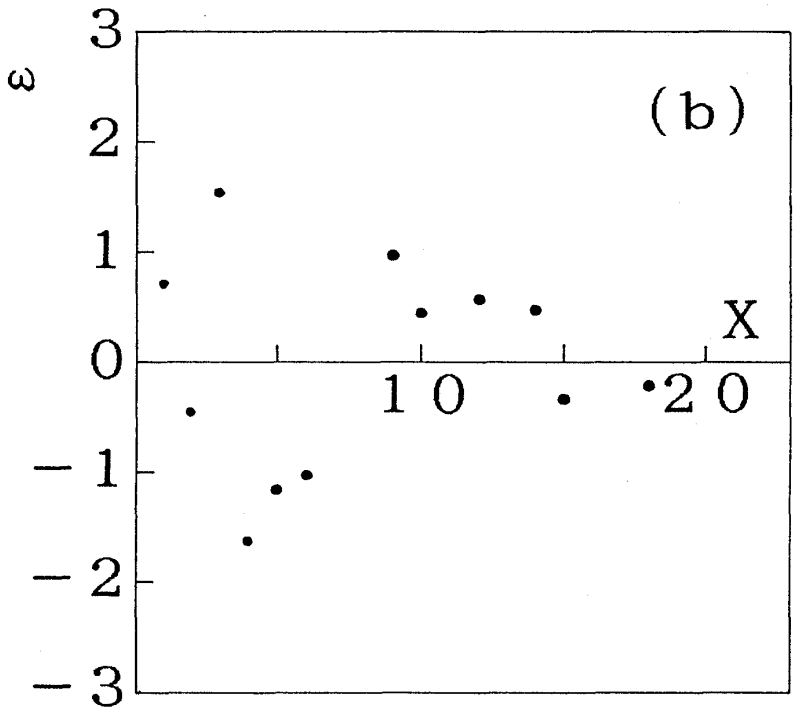
す．ここに標準化残差 ε_i (i=1−n) とは，残差 e_i をその標準偏差 s で割った無名数を指す (Chatterjee et al., 1981：5)．即ち

$$\varepsilon_i = e_i / s \tag{5}$$

$$e_i = v_i - \hat{v}_i \tag{6 a}$$



(a) 頭文字



(b) 尾文字

図2 対数分布（LG 型）に対する残差分析結果

$$s^2=Se/(n-2) \quad (6b)$$

$$Se=\sum_{i=1}^{n}e_i^2 \quad (7)$$

尚，nは縮約されたデータ数を表す（今の場合，n=18となる）．回帰モデルが妥当であるならば，$\varepsilon_i \sim N(0,1)$に従うはずである．ここに$N(\mu,\sigma^2)$は平均が$\mu$，分散が$\sigma^2$である正規分布を表す．$N(0,1)$分布の95%は$(-1.96,1.96)$の範囲にあるので，$\varepsilon_i$の約95%は$(-2,2)$の範囲に入ると期待してよい(Draper et al., 1968：90)．又，モデルが妥当ならば，残差プロットは特定の変動パターンをもたないはずである．図2(a)上のプロットを見ると，x=4（アルファベットのnに相当）に対して標準化残差が2を僅かに超えている．しかしながら，縮重点の存在を考慮すると，点の総数はローマ字アルファベットの数と同じ26個である為，この異常データを除いた全体の96%は$(-2,2)$の範囲に含まれる．更に，図2より，LG分布に対する残差プロットには，頭文字［図2(a)］，尾文字［図2(b)］共何等特徴的なパターンは認められないことが分かる．点集団から構成されるパターンのランダムネスを調べる為，ここでは次式で定義されるダービン・ワトソン比（ダービン・ワトソン統計量共言う）を考える．

$$d=Se^{-1}\sum_{i=1}^{n-1}(e_{i+1}-e_i)^2 \quad (8)$$

式(5)，(6)を用いると，この式は次の様に変形される．

$$d=(n-2)^{-1}\sum_{i=1}^{n-1}(\varepsilon_{i+1}-\varepsilon_i)^2 \quad (9)$$

dは，$0<d<4$の範囲内の値をとる．もし残差の系列がランダムであれば，$d\sim 2$，即ちdは2に近い値となり，隣接するデータ間に正の相関があれば0に，負の相関があれば4に近い値をとる（久米他，1987：180）．図2(a),(b)のデータに対してこの値を求めると，それぞれd=1.9, 2.1を得る[5]．よって，いずれの場合もdの値は2に極めて近く，当該残差系列はランダムであると判定される．今，比較の為，各端文字統計の直線分布（L型）に対して計算された残差プロットをそれぞれ図3(a), (b)に示す．これらを見ると，残差には明らかにU字形のパターンが存在し，点の集団がランダムに配置されているとは言い難い．実際，dの値を求めると，それぞれd=0.6, 0.7となった．これらはいずれも隣り合うデータ間に正の相関があることを意味している．結局，以上の解析結果より「ポイヤの表に見られる端文字統計はLG則に従っている」と結論することができよう．

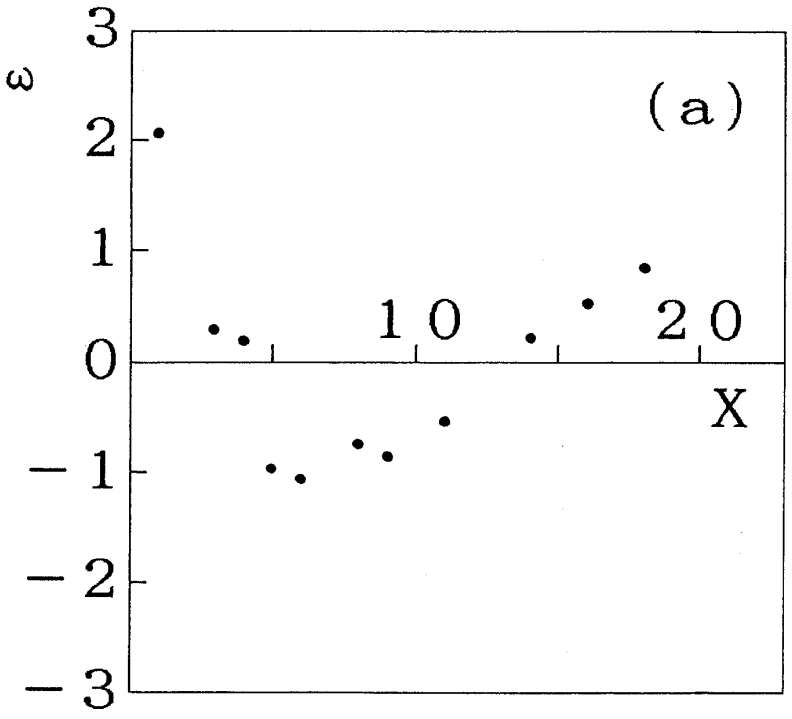
同様な特徴（LG型への適合性）は，英文一般において観測されている文字統計データ（今井，1984：96）を分析した場合にも認められた．このときの結果は

$$|r|=0.9850\ ;\ \hat{z}=12.73-8.75\log x \quad (10)$$

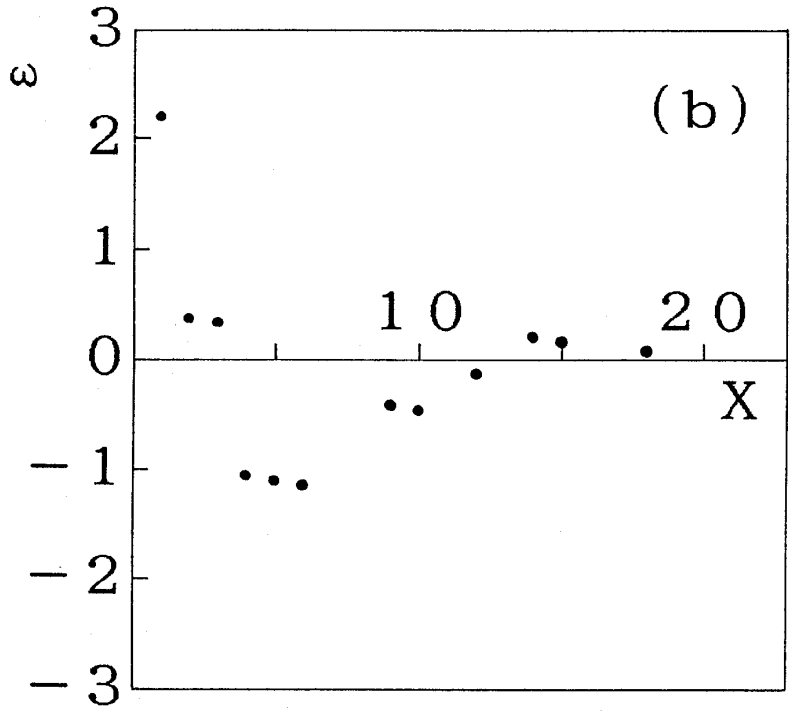
である．（他の分布型に対する|r|の値は以下の通り：EX：0.9341，ZP：0.7885，LN：0.8781，N：0.9786，L：0.9590.）一方，この文字統計に対して，シャノンの相対エントロピーは，h=0.89と評価されている（今井，1984：97）．比較の為，表2に示されているポイヤ表の文字頻度データに対して相対エントロピーの値を求めると

頭文字	h=0.90
頭の隣文字	h=0.81
尾文字	h=0.90
尾の隣文字	h=0.82

となった．この結果から，ポイヤの数詞表における文字統計に関して興味深い結論を引き出すことができる．即ち，端文字（頭と尾）については相対エントロピーが隣文字に比べて高く，その値は英文一般について計算されたものに匹敵するが，隣文字については相対



(a) 頭文字



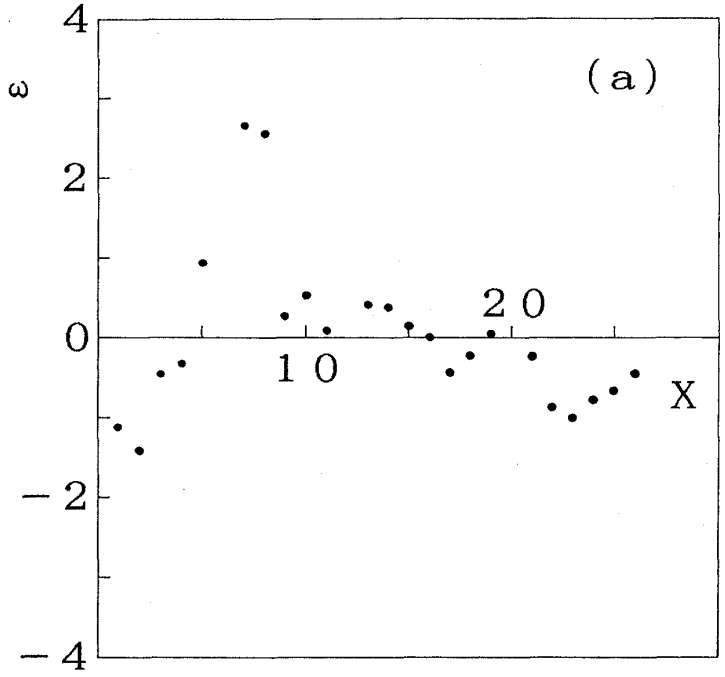
(b) 尾文字

図3 直線分布（L型）に対する残差分析結果

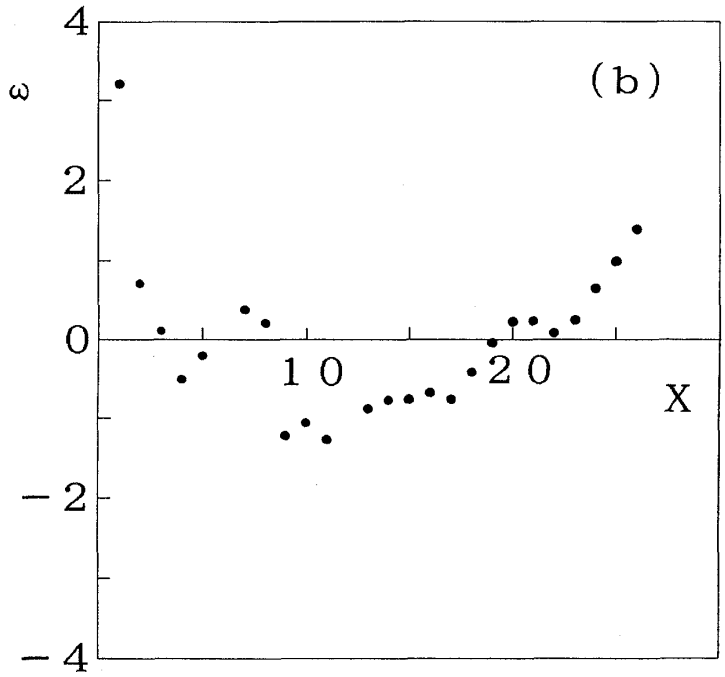
エントロピーがこれに比べて明らかに小さな値になっている．（この結果の説明については，次の段落で行う．）端文字の相対エントロピーについて認められた，この様な英文一般との類似性とは対照的に，残差パターンにおいては両者は顕著な違いを示した．即ち，英文一般の文字統計データに対して残差分析を行いｄの値を求めると，式(10)の分布（LG

型）に対してd＝0.55となった．この結果は隣接するデータの間に正の相関が見られることを意味している．（因に，L型に対してはd＝0.48となった．）図2，3との比較の為，英文一般に対する残差分析結果を図4に示す．LG型［図4(a)］，L型［図4(b)］いずれに対しても特徴的な残差変動パターンが認められる．よって，英文一般における文字統計に対しては「LG型への適合性」は認められるものの，本論文で定義した意味での「LG則」には従っていないと結論される．

筆者は，ポイヤ数詞表の端文字統計に対し



(a)　対数分布（LG型）



(b)　直線分布（L型）

図4　英文一般における文字の出現頻度に対する残差分析結果

て見出された統計則（LG則）を「競合とフラストレーション[6]を伴う自己組織系がある種の相転移を経た段階で観測される普遍的特徴の一つである」と理解しており，この描像を支持する数理モデル(トーナメントモデル)を既に構築している(早田，1997 a，b)[7]．それでは，この様な統計則は何故頭文字と尾文字だけについて見られ，これらに隣接する文字については見られなかったのであろうか？この問に答える為には当該サイトの文字統計のみに注目しては不十分であり，隣接する文字間の相関を考慮に入れなければならない．即ち，英文一般についてしばしば指摘されるｔとｈ，ｑとｕなどの事例と同様に，表１からも文字間の相関を読み取ることができる．例えば，語頭がｔの場合，次に来るものはｒである場合が多く見られる．又語尾がｎとなっているときは，その前の文字はｅとなることが多い．この様な制約が原因となって，出現頻度における文字間の「自由競争」とそれに起因する「緊張状態」が緩和され，その結果として分布がLG則に従わなくなるものと理解される．

3．漢数詞との比較分析

本節では，ポイヤの表（表１）に記された印欧語族系数詞の単語長に着目し，これらとは全く別種の言語である漢数詞（表４）との比較研究を行う．この研究の目的は，数の概念を表記・伝達する上での人類共通の性向について考察する為の糸口を提供することにある．

先ず，表１を基に，印欧語族系10言語における数詞長の度数分布を調べることにする．結果を図５の実線で示す．平均値，標準偏差はそれぞれ3.97，$1.45^{1/2}$と計算された．図５より，語長が３から５までの領域に全体の83%が集中していることがわかる．今，この度数分布を２項分布B(n, p)で近似することを考える．未知パラメータｎ，ｐは，２項分布の性質を用いると次の連立方程式の解として決定される．

表４ 漢数詞とその素画展開

i	漢数詞	素画展開	L_i
1	一	－	1
2	二	－－	2
3	三	－－－	3
4	四	－－－－－	5
5	五	－－－－	4
6	六	－－－－	4
7	七	－－	2
8	八	－－	2
9	九	－－	2
10	十	－－	2

注：L_i(i=1−10)はストリング長を表す．

$$np=3.97 \tag{11 a}$$

$$np(1-p)=1.45 \tag{11 b}$$

これらより，p=0.63，n=6.2～６を得る．故に，当該度数分布（図５実線）はB(6, 0.63)なる２項分布で近似可能である．比較の為，この分布を図５の上に破線で示す．破線は実線の特徴を大まかに再現しているが，両者は良く一致しているとは言い難い．図５の縦軸の相対頻度p(y)は語長ｙを生成する確率として解釈できる．以上のことから，図５の実線で示された語長分布の生成要因をベルヌーイ試行列として理解することには難があると言えよう．本結果の説明については今後の検討課題としたい．

以下では，表１に示された各言語の数詞の語長分布に着目し，漢数詞の「語長」分布との間の類似度について計量分析を行う．解析に先立ち，漢数詞の「語長」を定義する必要がある．周知の様に，漢字を構成する最も基本的な要素は素画である．例えば，表４の中に見られる「四」の画数は５であり，これは５個の素画から構成されていることを意味する．今，漢字を構成する全ての素画を表４の中欄に示す様に１次元的（紐状）に並べる．

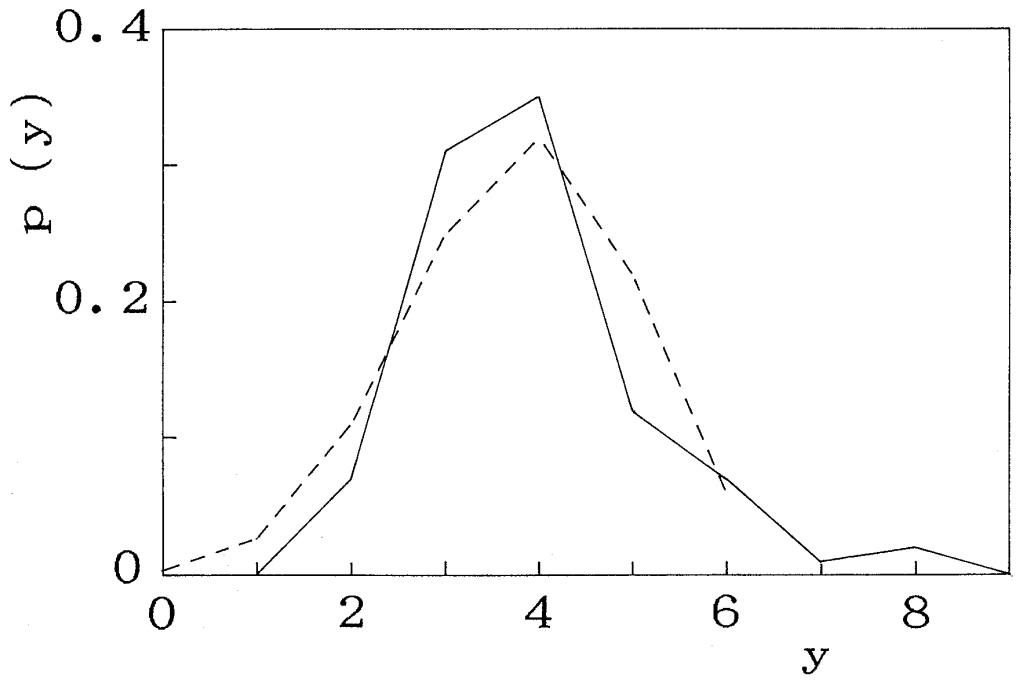


図5 ポイヤの表（表1）における相対頻度［p(y)］の文字長（y）依存性（実線）．破線はB(6,0.63)なる2項分布を表す．

ここでは便宜上この操作を「素画展開」と呼ぶことにする．生成された紐（ストリング）の長さは当該漢字の画数に一致し，これを「ストリング長（記号L_i）」と定義する．表4の右欄に各漢数詞のストリング長を明記した．

印欧語族系数詞と漢数詞の語長分布に着目したときの両者の比較分析結果を表5に記す．両数詞の語長分布の類似度を計量化する為，ヘリンガー距離D_H^2を採用した．これは非負値をとり，次式によって定義される（篠本，1992：68－70）．

$$D_H^2=\sum_{i=1}^{10}[p_i^{1/2}-q_i^{1/2}]^2 \qquad (12)$$

ここに$p_i=L_i/27$，$q_i=\lambda_i/\sum_{i=1}^{10}\lambda_i$であり，$\lambda_i$(i＝1－10)は表1に示す各言語の数詞の語長を表す（例：英語の場合，$\lambda_1=\lambda_2=\lambda_6=\lambda_{10}=3$，$\lambda_4=\lambda_5=\lambda_9=4$，$\lambda_3=\lambda_7=\lambda_8=5$）．考察の助けとして，表5にはこれらの他に，語長$\lambda_i$(i＝1－10)の特性値として，変動係数CV，平均値λ，中央値Me，最頻値Mo，範囲R，標準偏差sについても記載した．

表5より，D_H^2が最小となるのはデンマーク語に対してであり，フランス語，スペイン語，スエーデン語，イタリア語がこれに続いている．一般に，北欧系及びラテン系の言語において類似度が高い（D_H^2の値が小さい）傾向にあることが読み取れる．漢数詞とこれらの言語の数詞の語長分布に対して見られたこの様な類似は，単なる偶然によるものなのであろうか？ この問を検証する為，統計的仮説検定を行うことにする．検定を行うに当たり，帰無仮説Hを以下の様に設定する．

H：デンマーク語，フランス語，スペイン語，スエーデン語，イタリア語の各数詞と漢数詞の間に認められた語長分布の類似性は偶発的なものである．

本仮説を検定する為，$2\leq\lambda_i\leq 8$という制約の下で多数の数詞を無作為に生成し，生成された各数詞と漢数詞の間の語長分布の類似度を計算した．無作為抽出には乱数（岡本他，1977：184）を使用した．この様に生成された100個の仮想言語に対して計算されたヘリンガー距離D_H^2の度数分布を図6に示す．得ら

表5 ポイヤの表（表1）と漢数詞（表4）の比較分析結果

	英語	典語	デ語	蘭語	独語	仏語	西語	伊語	ポ語	ハ語
D_H^2	0.052	0.034	0.025	0.056	0.045	0.028	0.029	0.037	0.076	0.086
CV	0.213	0.174	0.287	0.158	0.178	0.267	0.209	0.314	0.289	0.313
λ	3.9	3.1	2.9	4.0	4.2	3.9	4.3	4.3	5.4	3.9
Me	4	3	3	4	4	4	4	3	5	3.5
Mo	3	3	2	4	4	4	4	4	5	3
R	2	2	2	2	3	4	3	4	5	4
s	0.831	0.539	0.831	0.632	0.748	1.04	0.900	1.35	1.56	1.22

注：漢数詞データの特性値は，表4より，CV＝0.441，L＝2.7，Me＝Mo＝2，R＝4，s＝1.19

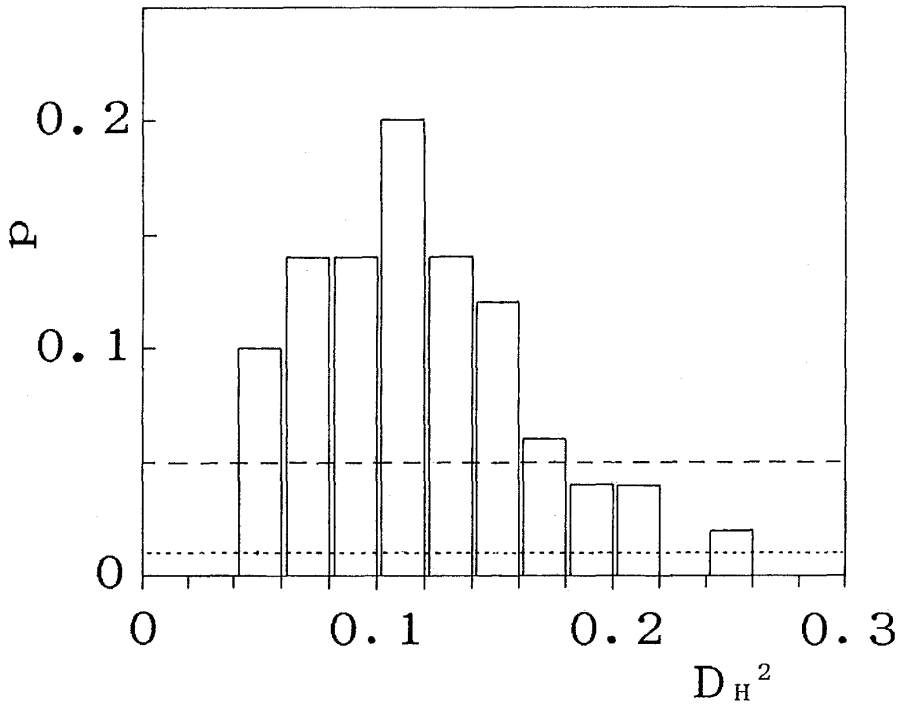


図6 無作為に生成された100個の仮想言語に対するヘリンガー距離のヒストグラム（△ D_H^2＝0.02）．p，D_H^2 はそれぞれ生成確率，当該言語の数詞と漢数詞とのヘリンガー距離を表す．D_H^2 の平均値，中央値，最頻値はそれぞれ0.117，0.111，0.111である．図中に描かれている破線，点線はそれぞれ5％，1％有意水準を示す．

れた分布は D_H^2＝0.111近傍にピークが存在し，右方向に歪んだ形をしている．図中に描かれている破線，点線はそれぞれ5％，1％有意水準を示している．図6より，$D_H^2 \leq 0.04$ の領域でp＝0となっていることがわかる．この結果と表5に示した D_H^2 の計算結果から，帰無仮説Hは棄却される．よって「本検定は有意水準1％で有意である」と結論することができる．

上で見た語長分布の類似性が偶然に依るものでないとすると，それでは一体この類似性の起源を何処に見出すことができるのであろうか？ 筆者はこの類似性を，かつて世界各地で使用されていたであろうと推定される4進法の名残（加藤，1996：92－93,167－168）に求めることができるのではないかと考えている．例えば，古代エジプトで使用された長さの単位に〝palm〟並びに〝digit〟というものがある．1 palmは掌の横幅に相当し，1 palm＝4 digitという関係がある．これは，掌

の幅の中に，人差し指から小指までの4本の指があることに起因していると思われる．又，印欧祖語，ラテン語，リトアニア語の数詞5は4が区切りになって構成されている．即ち，印欧祖語の数詞5は〝penque〟であるが，これは〝-pen-que〟と分解される．ここに〝pen〟は現代英語の〝one〟を，〝que〟は〝hand〟に対応する．〝hand〟は親指を除いた4指を表す．一方，我が国の江戸時代の貨幣単位に目を向けると，「1分＝4朱」「1両＝4分」という様な4進法の構成をもつものが見られる（加藤，1996：92－93）．さて，漢数詞「四」の初文は，籀文の字形が示す様に四横画を重ねた形で示されていたという．即ち，かなり古い時代には，「二」と「二」を上下に配置することによって4を表した（白川，1996：639）．又，「五」という漢字は，斜めに交錯する木を以て作られた器物の蓋の形を仮借したものである．即ち，卜文では一より四までは横画を重ね，五に至って交画とした（白川，1996：478）．

4．むすび

ポイヤの数詞表を構成する100個の印欧系数詞に対して，頭文字，尾文字並びにそれらに隣接する文字の出現頻度について回帰分析を行い，単語中の文字サイトと統計則の関係について新知見を得た．引き続き，この表に示された各言語の数詞の語長分布に着目し，漢数詞のストリング長（画数）分布との類似度について計量分析を行った．その結果，北欧系及びラテン系の言語において高い類似度を見い出した．類似度を判別する為の指標として，ヘリンガー距離を採用した．更に，得られた計算結果に対して統計的仮説検定を行い，この様な類似性は偶然に依るものではないという結論を導いた．言語は人類による森羅万象の様々な「理解の仕方」を反映していると考えられている（箕浦，1993；中島，1995）．よってこの発見は，数の概念を表記・伝達する上での人類共通の性向について哲学的考察を行う際の糸口を提供するものである(8)．

注

(1) ポイヤの方法以外の計量言語学的研究については，例えば安本美典による解説（安本，1995）を参照されたい．

(2) 術語・人名の表記法は辞典（日本数学会編，1985：1586）に倣った．

(3) 対数分布（LG型）と直線分布（L型）以外は全て目的変数yは正でなければならない．従って，実際の回帰分析では，これら2つの分布への回帰に際してはy＝0の点［頭文字の場合：(x, y)＝(18, 0)］を含めたが，他の4つの分布への回帰に際してはy＝0の点は除いた．

(4) これらの変換式は，いわゆるBox-Cox変換（冪変換共言う）を一般化したものである．

(5) 「頭の隣文字」「尾の隣文字」に対して同様の計算を行うと，それぞれd＝1.30, 1.54を得る．

(6) 「フラストレーション」という術語は元々精神分析学の文脈で開発されたものである．フロイトは，フラストレーションを何らかの妨害によって要求の充足が阻止されることであるとし，行動のメカニズムを説明する為の仮説的概念としてこの術語を使用した．しかしながら，要求の充足の阻止によって有機体内に生じた「緊張状態」を指す用語として，又「葛藤状態」と同義に使用されることもあり，定義は一意的ではない（坂田他，1996：406）．1977年Toulouseは，スピングラスの挙動を説明する為のモデルとして，統計物理の分野にこの概念を類比的に導入した．即ち「幾つかの相互作用が競合している為，それら全ての相互作用についてその効果を最大限に発揮できないような状況（系）」をフラストレーション（フラストレートしている系）と呼んだ（東京大学物性研究所編，1996：842）．スピン系に限らず，この種の相互作用の競合の問題は，現代物理の様々な場面において現れる．物性物理学者・高山一

は，この概念の数学や工学の分野への一般化を試みている(同上，1996：842)．一方，Chowdhuryは，フラストレーション描像の社会や生命進化への応用可能性を示唆している(Chowdhury, 1986：275－276)．

(7) イスラエルの物理学者Kanter等は，2パラメータ・ランダムマルコフ過程によって，本統計則の説明を試みている(Kanter et al., 1995)．

(8) 数の概念の文化による相違については，加藤良作による考察がある(加藤，1996：165－174)．

参考文献

Chatterjee, S. and Price, B. (1981)『回帰分析の実際』佐和隆光，加納 悟訳，新曜社

Chowdhury, D. (1986) Spin Glasses and Other Frustrated Systems, World Scientific, Singapore

Draper, N. R. and Smith, H. (1968)『応用回帰分析』中村慶一訳，森北出版

早田和弥(1997 a)「現代社会にはびこる不正・不合理の統計分析――社会情報学研究の視点から――」『社会情報学研究』No.1：101-111

早田和弥(1997 b)「歴史研究への非線形動力学的アプローチ」『文理シナジー学会大会研究一般発表要旨集』講演番号6

今井秀樹(1984)『情報理論』昭晃堂

Kanter, I. and Kessler, D. A. (1995) Markov processes: linguistics and Zipf's law, Phys. Rev. Lett., Vol.74, No.22: 4559-4562

加藤良作(1996)『数詞って何だろう』ダイヤモンド社

久米 均，飯塚悦功(1987)『回帰分析』岩波書店

箕浦信勝(1993)「<クラウス博士講演報告>言語絶滅の危機目前に迫る――我々言語学者は何をなすべきか」『月刊言語』Vol.22, No.8：12-17

中島 久(1995)「スワヒリ語世界へ」『学士会会報』No.808：73-77

日本数学会編(1985)『数学辞典(第3版)』岩波書店

岡本雅典，鈴木義一郎，杉山高一(1977)『基本統計学』実教出版

坂田成輝，浜口晴彦他編(1996)『現代エイジング辞典』早稲田大学出版部

篠本 滋(1992)『情報の統計力学』丸善

白川 静(1996)『字通』平凡社

東京大学物性研究所編(1996)『物性科学事典』東京書籍

安本美典(1995)『言語の科学』朝倉書店

1998年1月12日受付
1998年2月23日受理