文字の統計：グラフィック端末による分析

文字の統計

——グラフィック端末による分析——

田 中 卓 史

1. はじめに

近年，電子計算機を利用した事務処理の分野で，住所，氏名，会社名などの漢字データを扱うことの必要性から，漢字入出力機器の開発が盛んになっている。漢字を便利に扱える機器の出現は，同時に漢字仮名まじり文の計算機処理という新しい情報処理の分野を開くものである。しかし，これらの漢字機器も入力に関しては十分便利なものであるとは言い難い。これは本質的には日本語を書き表す文字，特に漢字の数の多さに原因があるが，またできるだけ多くの文字を文字盤面に収容しようとする設計方針にも問題があるように思える。

単に計算機入力の点から見れば，日本語を書き表す漢字は少ないほど都合が良いのであるが，一方，漢字は漢字仮名まじり文において (1)文節の切れ目を示す，(2)語のまとまりを示す，(3)同音異語を区別する，など文を読みやすくするための重要な役割を演じていることが知られている。漢字仮名まじり文の計算機処理を行うに際し，効率の良いシステムを作るためには，日本語を書き表す文字についての基礎的なデータを集めておくこと，特に漢字の文中における役割や性質について十分把握しておくことが必要である。

ここでは新しく導入した東大大型計算機センターの TSS 端末装置（Tektronix 4006，グラフィック端末）を用いて漢字仮名まじり文を構成する文字を統計的に調べ，文字の機能を明らかにすることを試みる。この装置は計算機からの出力結果を直接グラフの形で表示することができるので，集計や分析において単に数値の列をながめているだけでは見出すことのできなかった文字の集団としての性質を視覚的に明らかにすることができる。

調査の対象とした漢字仮名まじり文は，現在，計算機入力が進められている

「高校教科書の用字用語調査」のデータの内，教科書のページを単位として1/20にサンプリングされたパイロットデータである。ページ単位のサンプリングのため特定の文字や語が片寄って含まれている可能性があるが，9教科（日本史，世界史，政治経済，倫理社会，地理，物理，化学，生物，地学）を対象としているのでこの片寄りはいく分緩和されていると思われる。

2. 教科別の文字数

調査データには延べ48096個の文字が出現し，それらは1525個の異なった文字（盤外特殊記号は区別せず1種類と数えている）から構成されている。

図1は延べ文字数の教科別，字種別の内訳を表したもので，x軸（横軸）は教科別の割合を多い順に示し，y軸（縦軸）は各教科ごとに字種別の割合を示している。グラフ中の英文字は次の教科名を示す。

J：日本史	W：世界史	E：政治経済
M：倫理社会	A：地理	P：物理
B：生物	C：化学	G：地学

字種は次の9種類に分け，x軸に近い方から順に領域を割り当てている。

1：漢字	2：平仮名	3：片仮名
4：英字	5：数字	6：7，8，9以外の記号
7：盤外特殊記号	8：ピリオド（句点）	9：コンマ（読点）

字種の構成には各教科ごとの特徴が見られる。これを見やすくするために，図1を教科ごと・字種ごとの長方形に切り離して，字種ごとにy軸の大きさ（教科別字種の割合）に従って並べると，図2，図3が得られる。

図2は漢字，平仮名，片仮名に関して画いたものである。漢字(グラフ中央)を多く含む教科は日本史，政治経済，地理，…の順となり，社会科には漢字が多く理科には少ない傾向が見られる。平仮名（グラフ上方）を多く含む教科は生物，倫理社会，物理，…の順となり，地理，世界史，化学は少なくなっている。

片仮名を多く含む教科は地理，世界史，化学，…の順となり，平仮名の少ない分だけ片仮名（グラフ下方）が増えていたことになる。

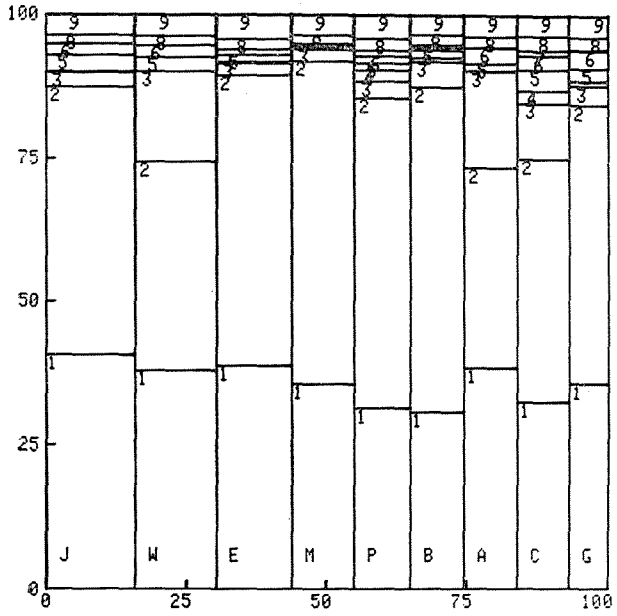

図1　文字の教科別，字種別割合（延べ文字数）

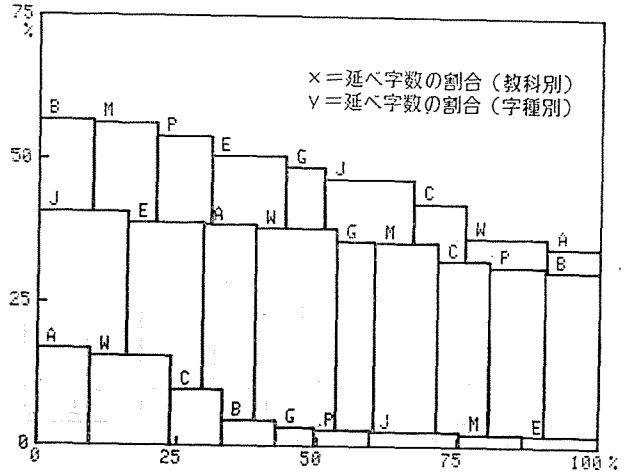

図2　漢字（中），平仮名（上），片仮名（下）の教科別，字種別割合

図3は英字，洋数字，記号類（6～9），に関して画いたもので，y軸の長さを5倍に拡大している。記号類を示すグラフ（上方）において，点線で示した部分まではピリオド（句点）とコンマ（読点）で占められている。点線よりも上の部分をそれ以外の記号が占めている。点線の部分に着目すると，句読点は生物以外の理科に多く，政経以外の社会科に少なくなっている。一般的に社会科よりも理科の方が簡結な文が用いられている様子がうかがえる。

図4は全教科のデータに関して字種別延べ文字数の割合を，直観的把握ができるように矩形の面積で表したものである。もし，文字種の使用頻度に比例した文字入力装置の文字盤面を作るとするならば，このような割合になろう。（このグラフは領域の大きい順に長方形の長い方の辺から必要な領域を切り取るアルゴリズムで画いている。）

図5は教科別に異なり文字数の割合を示したものである。x軸は教科別の異なり文字数に比例した大きさとなっているが，大多数の文字が教科間で重複するため，各教科の異なり文字数を加えたものが全教科異なり文字数（1525字）

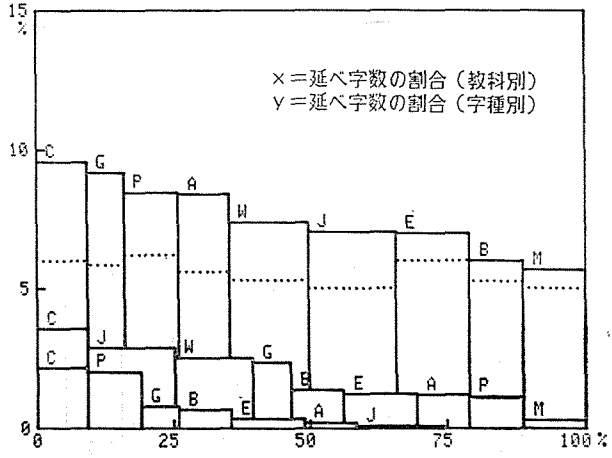

図3　英字（下），洋数字（中），記号類（上）の教科別，字種別割合

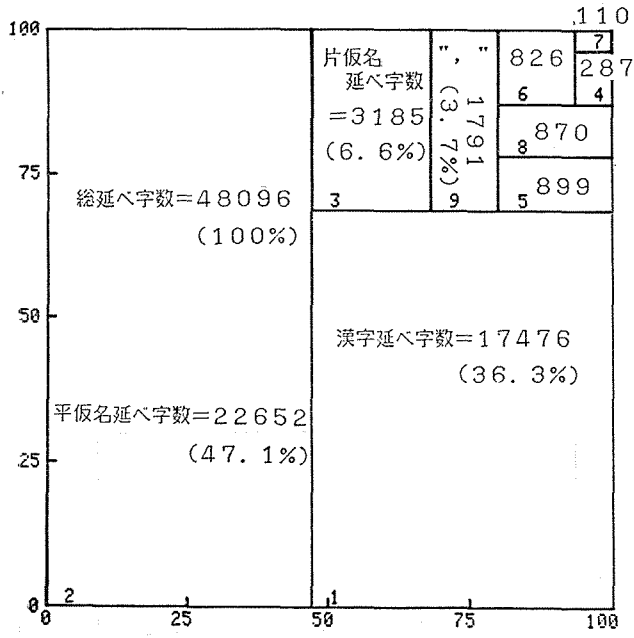

図4　延べ文字数の字種別割合

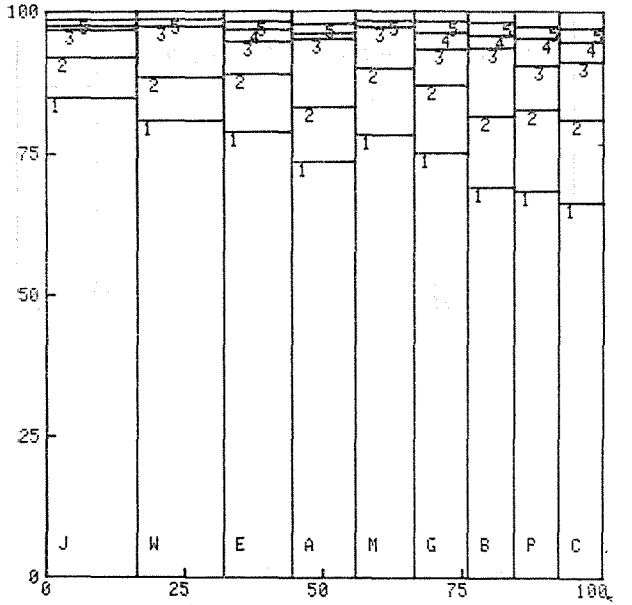

図5　異なり文字数の教科別，字種別割合

ではない。この場合に x 軸の 100 に相当する値は，各教科の異なり文字数を単純に加えた値（5343字）になる。一般的にみて，データ数の多い教科ほど，漢字の占める割合が多くなり，平仮名の占める割合が少なくなる傾向を示している。平仮名は教科書の1/20サンプリング程度の量のデータでも，そのかなりの部分が出現するのに対し，漢字は各教科の 1/20 サンプリング程度のデータでは，十分飽和するまで出現していないことに起因する。

図6は全教科の異なり文字数に，字種の占める割合を図4と同じように画いたものである。この図は各字種の示す矩形の面積がその文字数に比例するので，従来の文字入力装置の文字盤面に近いものである。

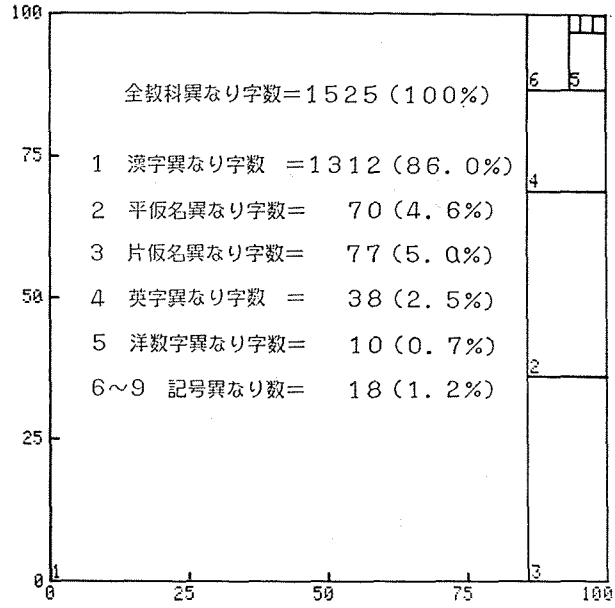

図6 異なり文字数の字種別割合

3. 文字の出現頻度

(1) 文字の順位

図7は各々の文字を出現頻度順に並らべ，順位と頻度との関係を表したものである。出現頻度が10数回以下の部分では同一頻度の文字が多くなるので，グラフを x 軸方向に延びる線分の集まりとして表している。一つの線分に含まれ

る文字は左端の点が順位を示し，線分の長さが頻度の等しい文字の数を示す。

y軸の値を延べ文字数(48096)で割ると，従来よく用いられてきた使用率が求まる。上位の文字については表1に順位と頻度の関係を示している。

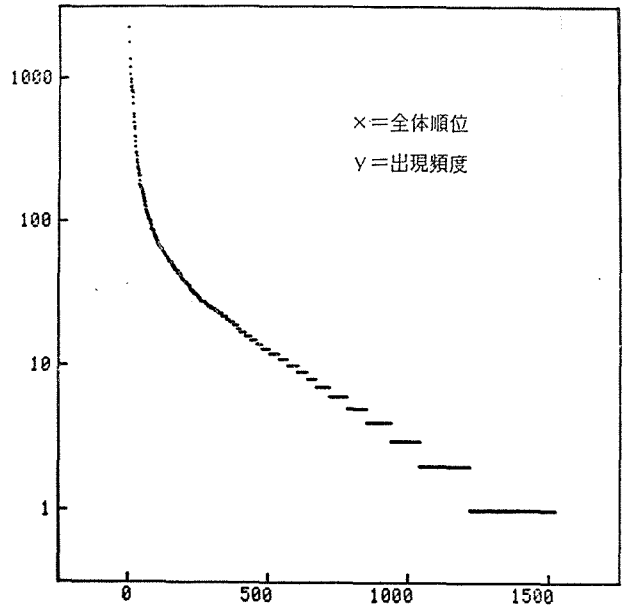

図7　文字の順位と出現頻度

図8はx軸に全体順位を，y軸に字種別の順位を表している。図中の数字（1～5）は前述の字種を示す。字種ごとに異なった曲線が画かれている。最初に平仮名(2)が現れるが50位付近で飽和の傾向を示す。次いで32位から片仮名(3)が現われ，徐々に増え続け600位付近で平仮名の数を追い越し，次第に飽和の傾向を示す。これは片仮名が世界史や地理において外国の人名や地名などを表す表音記号として用いられることが多いためで，使用頻度は低いものの，平仮名よりも多くの濁音，半濁音，幼音を表す文字が用いられることによる。

漢字(1)は33位から現れ始め，全体順位とともにほぼ傾きが1で直線的に上昇する。これは5～60位以上の高順位では漢字が支配的となり，間に他の字種の文字がわずかに混じる状態になっていることを意味する。

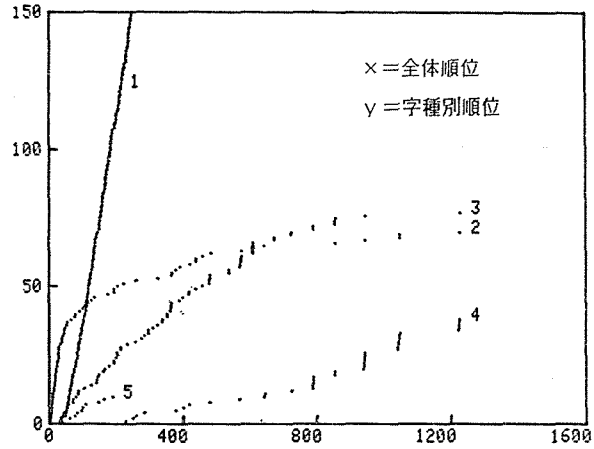

図８　文字の全体順位と字種別順位

図９は字種別順位と出現頻度の関係を示したものである。平仮名(２)もすべての文字が高頻度で用いられるのではなく，出現頻度が１回になるまで分布していることがわかる。洋数字(５)は傾きが大きく，各々の数字（0，1，2，…，

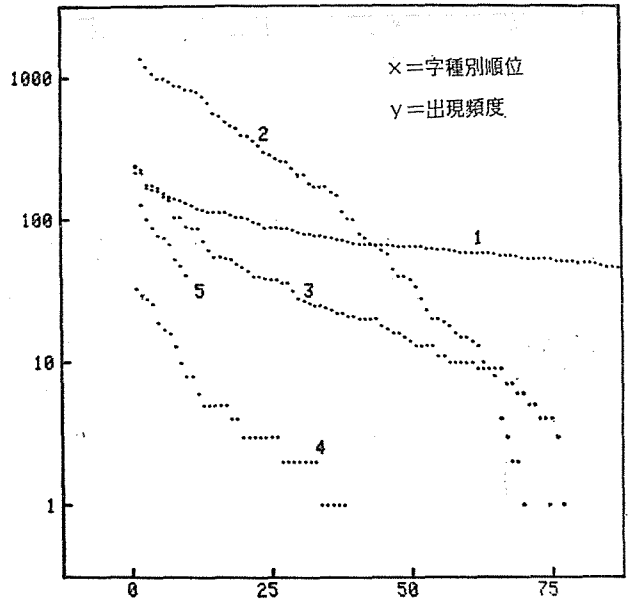

図９　字種別順位と出現頻度

9)の間で使用頻度が大きく異なっている。これは洋数字が世界史や日本史などで年号に多く用いられた影響によるものと思われる。表1に字種別に上位20位までの文字を示す。

表1　文字の字種別順位と全体順位および出現頻度

字種別順位	漢字			平仮名			片仮名			英字			洋数字			記号他		
		順位	頻度		順位	頻度		順位	頻度		順位	頻度		順位	頻度		順位	頻度
1	国	33	238	の	1	2268	ア	32	244	A	227	33	1	37	217	,	2	1791
2	地	36	218	に	3	1369	ン	35	228	B	250	30	2	61	128	.	11	870
3	化	42	179	る	4	1209	イ	48	196	C	259	28	3	76	102	・	26	308
4	大	43	177	と	5	1076	リ	50	165	m	284	26	0	83	89	)	40	194
5	生	48	169	を	6	994	ス	51	161	a	377	19	4	96	78	(	43	170
6	人	55	148	は	6	994	ル	53	156	c	399	17	9	100	75	盤外	70	110
7	動	56	146	た	8	939	ー	58	138	H	418	16	5	109	68	—	227	33
8	中	57	142	て	9	887	ラ	72	105	p	478	13	8	147	53	⌐	323	23
9	的	58	138	が	10	871	カ	72	105	b	570	10	6	163	48	¬	336	22
10	物	60	132	い	12	830	シ	82	90	O	642	8	7	190	41	～	399	17
11	業	62	126	な	13	819	ド	87	87	g	642	8	—	—	—	%	570	10
12	分	63	120	し	14	795	ト	87	87	N	724	6	—	—	—	=	607	9
13	方	64	117	で	15	740	ジ	105	71	S	787	5	—	—	—	／	607	9
14	一	65	115	れ	16	671	フ	125	62	n	787	5	—	—	—	〔	642	8
15	発	65	115	こ	17	568	ロ	142	55	l	787	5	—	—	—	〕	642	8
16	民	68	113	か	18	539	ム	142	55	k	787	5	—	—	—	+	724	6
17	水	68	113	ら	19	494	ク	145	54	f	787	5	—	—	—	『	1221	1
18	年	71	108	っ	20	459	ウ	147	53	g	854	4	—	—	—	』	1222	1
19	本	72	105	も	21	445	エ	161	49	d	854	4	—	—	—	—	—	—
20	会	75	104	す	22	396	ギ	165	47	T	940	3	—	—	—	—	—	—

(2)　順位と延べ文字数

図10は文字の全体順位に対し，各字種ごとに累積延べ文字数を求め，全延べ文字数に占める割合を表したものである。最終順位では図4に示した字種別延べ文字数の比率に一致する。平仮名(2)はごく若い順位までの間に，延べ文字の大部分を占めてしまう様子がわかる。平仮名(2)と片仮名(3)は異なり文字数では大差ないが，このグラフでは全く異なった様子を示している。

図11は字種別の順位とその順位までの延べ文字数の和(累積延べ文字数)の関係を表わしている。漢字は最終的に，1312位において延べ17476文字となる。

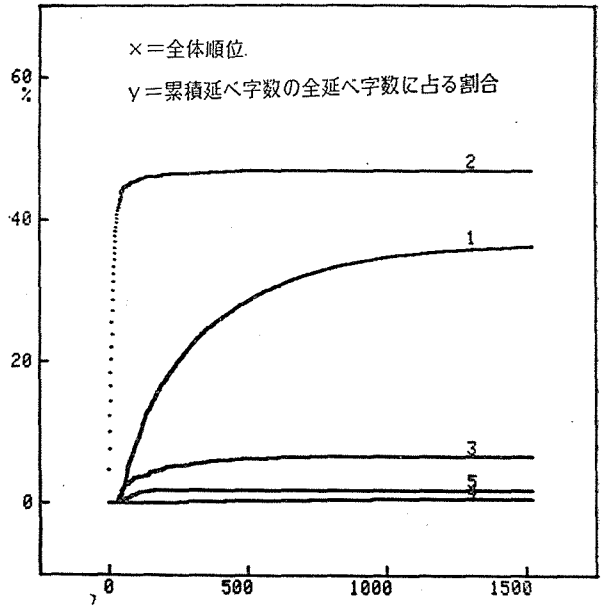

図10　全体順位と字種別累積延べ文字数の割合

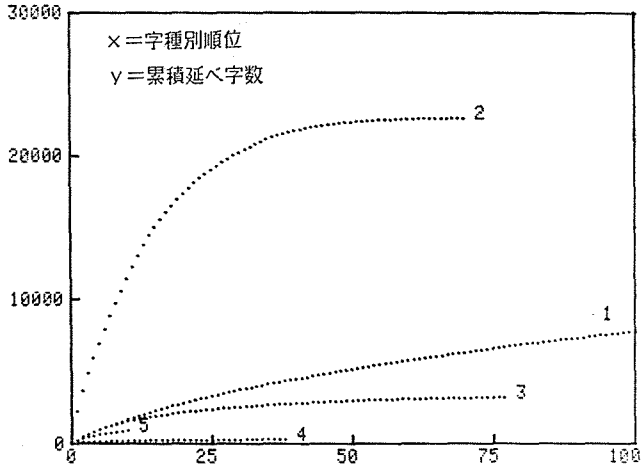

図11　字種別順位と累積延べ文字数

図12は字種別順位と字種別の延べ文字数の累積比率を示したものである。平仮名(2)も片仮名(3)も，わずか10位までの文字で延べ文字数の50%を越え

ていることがわかる。平仮名と片仮名とでは 10 位以上の部分でグラフの形がかなり異なっている。これは図9において平仮名と片仮名の傾きの違いに見られることと本質的に同じで，平仮名の方が若い順位の文字が集中的に使用される傾向にあることを示している。

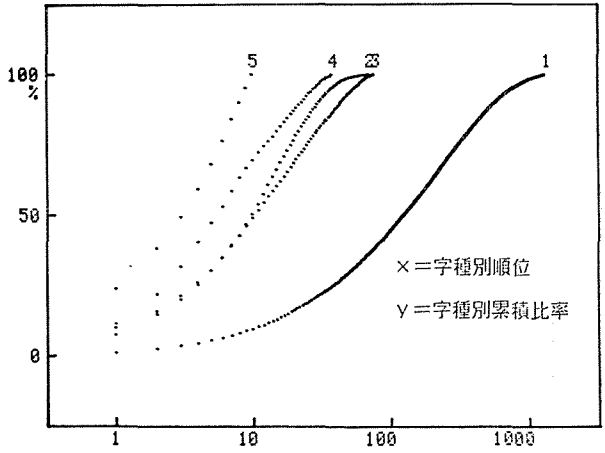

図12　字種別順位と字種別累積比率

4.　文字と語

(1)　文字の用いられる異なり語の数

漢字の中には同程度の使用頻度でも多くの異なった語に用いられるものと，特定の語にしか用いられないものがある。田中章夫[2]は一つの漢字がどれだけの語の表記に関係するかを明らかにするための統計的尺度として，いくつかのカバー率の概念を提唱している。この中の異なり語カバー率は次式で定義されている。

$$\text{異なり語カバー率}_{(‰)} = \frac{\text{ある漢字の用いられた異なり語数}}{\text{異なり語総量}} \times 1000$$

この考えは文字全体に拡張することができる。分母の異なり語数は調査対象に関して定まる定数なので，分子にのみ着目し，個々の文字が幾つの異なった語（M単位[3]）に用いられるかについて調べる。

一つの調査対象となった文字の集団において，各々の文字に「頻度」という値が定まるように，「異なり語数」もまた文字ごとに定まる値である。図7～図12を画くプログラムで，文字の「頻度」を「異なり語数」で置換えると，同様なグラフを画くことができる。

図13は図7に対応するもので，文字の用いられる異なり語数を求め，多い順に並べたものである。最も多くの異なり語数を持つものは平仮名の「る」で，148 個の異なった語に用いられている。次いで「い」，「か」，「っ」，「し」，「ン」，「く」，「ル」，…と続く。（表2参照）

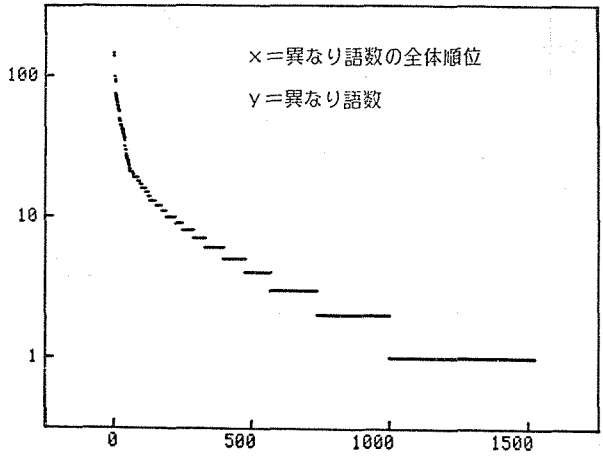

図13　文字の用いられる異なり語数とその順位

図14は図8に対応するもので，文字の異なり語数の全体順位と字種別順位の関係を示している。漢字(1)，平仮名(2)，片仮名(3)，英字(4)以外の字種は，M単位語では単独で1単位と見なされ，異なり語数が1となるので，このグラフには表していない。

図15は図9に対応するもので，字種別に異なり語数の順位と異なり語数の関係を示している。図15が図9と基本的に似た形をしているのは，各文字の頻度と異なり語数が完全に独立にとりうる値ではないこと，すなわち，

異なり語数 ≦ 出現頻度

の制限が効いていると考えられる。

表2には異なり語数の字種別順位が上位20位までのものを全体順位，異なり語数とともに示す。（同じ異なり語数を持つものは頻度順に示す）

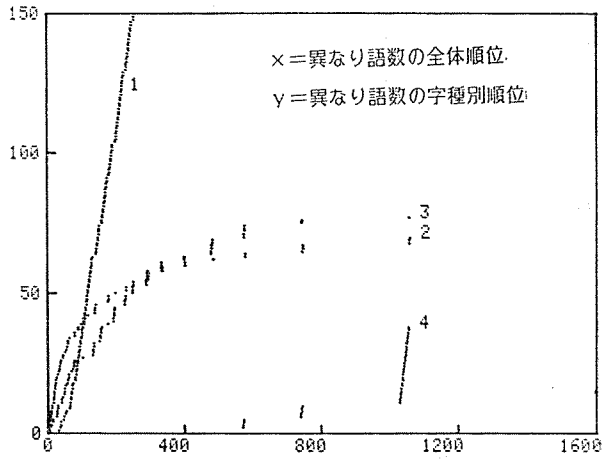

図14　異なり語数の全体順位と字種別順位

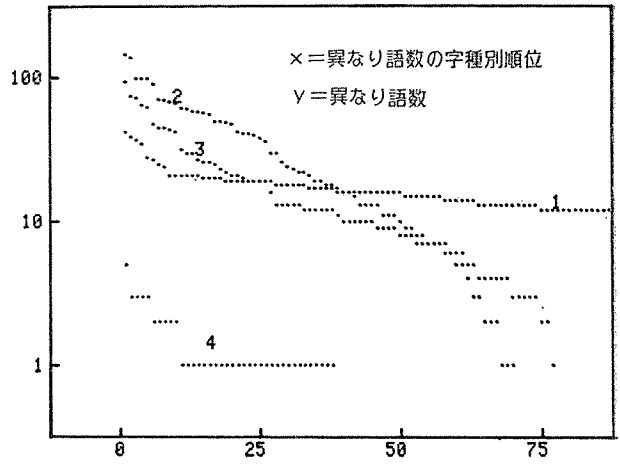

図15　異なり語数と字種別順位

表2　文字の異なり語数とその順位

字種別	漢字			平仮名			片仮名			英字		
順位		全体順位	異なり語数		全体順位	異なり語数		全体順位	異なり語数		全体順位	異なり語数
1	地	30	42	る	1	148	ン	6	94	a	400	5
2	国	36	39	い	2	140	ル	8	75	C	572	3
3	大	38	37	か	3	100	ー	9	73	c	572	3
4	分	40	35	っ	3	100	ス	14	65	l	572	3
5	一	46	28	し	5	99	ア	15	63	n	572	3
6	定	47	27	く	7	91	ラ	25	48	m	741	2
7	人	52	25	り	10	71	イ	27	45	N	741	2
8	水	54	24	ら	11	70	リ	27	45	o	741	2
9	動	61	21	た	12	69	シ	29	44	s	741	2
10	中	61	21	な	13	67	ト	30	42	h	741	2
11	体	61	21	ま	16	62	ク	41	32	A	1002	1
12	実	61	21	す	17	61	カ	42	30	B	1002	1
13	学	61	21	わ	18	59	ド	42	30	H	1002	1
14	成	61	21	き	19	58	マ	47	27	p	1002	1
15	生	70	20	れ	20	57	ロ	49	26	b	1002	1
16	方	70	20	え	21	56	ツ	49	26	g	1002	1
17	代	70	20	う	22	50	タ	52	25	O	1002	1
18	制	70	20	め	22	50	ジ	56	23	f	1002	1
19	物	75	19	つ	24	49	フ	58	22	k	1002	1
20	子	75	19	さ	25	48	ナ	61	21	S	1002	1

異なり語数の多い漢字「地」，「国」がどのような語に用いられていたが KLIC（文字を単位とするKWIC）を用いて調べると次のようになる。括弧の中は出現頻度を示す。

地（218）

地域（30），地方（30），地（30），土地（18），地頭（17），地図（13），地球（10），地形（7），地震（7），地位（5），地上（4），地中（4），山地（4），盆地（4），地質（3），地帯（3），各地（3），耕地（3），地理（2），本地（2），要地（2），地面（1），地勢（1），地下（1），地表（1），地類（1），地狭（1），地点（1），地かく（1），聖地（1），台地（1），局地（1），田地（1），境地（1），現地（1），産地（1），分地（1），立地（1），測地（1），陸地（1），加地子（1）

国 (238)

国民 (29)，国 (70)，中国 (18)，国家 (16)，帝国 (10)，国際 (9)，国内 (8)，外国 (8)，諸国 (8)，各国 (8)，王国 (5)，国交 (3)，国王 (3)，全国 (3)，開国 (3)，大国 (3)，一国 (3)，国政 (2)，国司 (2)，国有 (2)，国立 (2)，建国 (2)，清国 (2)，他国 (2)，鎖国 (2)，本国 (2)，国産 (1)，国土 (1)，国務 (1)，国力 (1)，国債 (1)，国策 (1)，国境 (1)，帰国 (1)，同国 (1)，亡国 (1)，小国 (1)，両国 (1)，報国 (1)

図16は図10に対応するもので，x軸は異なり語数の全体順位を示し，y軸は，各字種ごとにその順位までの各文字の異なり語数の総和を求め，文字全体についての異なり語数の総和に占る割合を表したものである。各々の文字を含む語は文字間において重複する可能性があるので，真の異なり語数よりも大きな値となっている。図10と比較すると延べ文字数の多かった平仮名が伸びずに，漢字が大きく延び，最終的には全体の 60% 近くを占めることになる。漢字と仮名とでは異なり文字数が大きく違うことからの当然の結果であるが，漢字から構成される語の方が仮名から構成される語よりはるかに種類が豊富である様子がうかがえる。

図17は図11に対応するもので異なり語数の字種別順位に対して，字種別にその順位までの異なり語数の総和を示している（真の異なり語数よりも文字間の語の重複の分だけ多くなる)。図10と比較すると片仮名のグラフが上方に大きく移動していることがわかる。これは片仮名が外国の人名や地名などの固有名詞に用いられるためで，頻度のわりに異なり語数が多くなること，また文字数の多い語からなるため，文字ごとに同一の語が何度も重複して数えられることによる。

図18は図12に対応するもので，x軸は異なり語数の字種別順位，y軸は字種別異なり語数の累積比率を表す。図12よりも全体的に右側に移動している。

これは，図9よりも図15のグラフの方が傾斜がゆるやかになっていることに起因し，若い順位の文字が頻度の場合ほどは，異なり語の全体に影響を与えてないことを意味する。

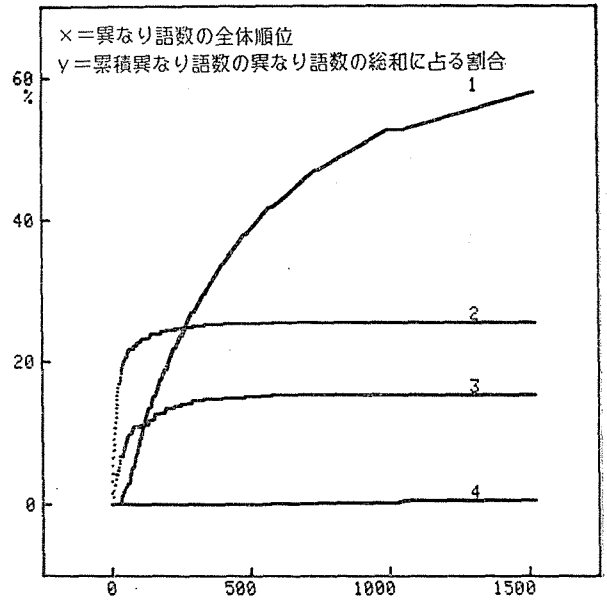

図16　異なり語数の全体順位に対する字種別累積異なり語数の異り語数の総和に占める割合

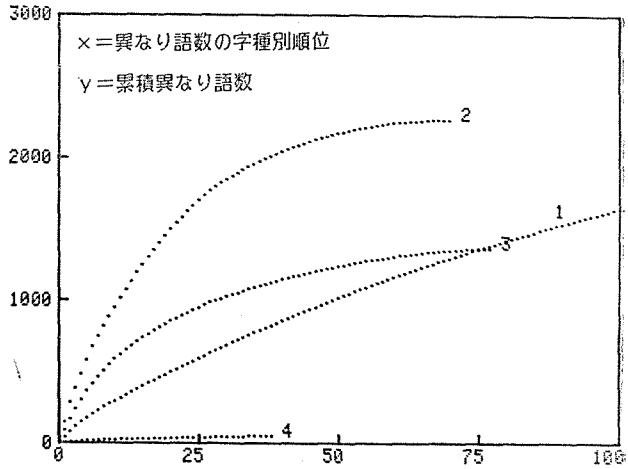

図17　異なり語数の字種別順位と字種別累積異なり語数

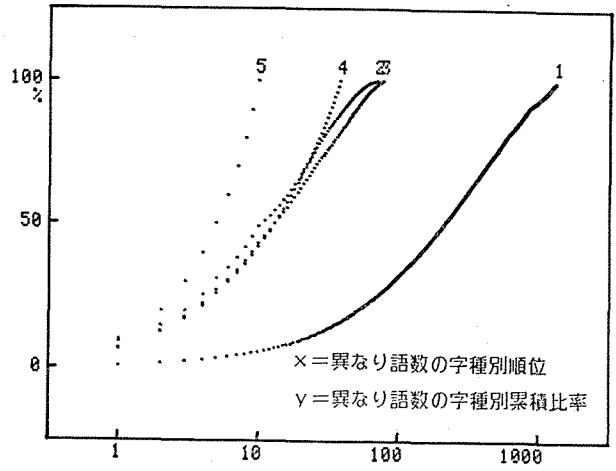

図18　異なり語数の字種別順位と累積比率

(2)　文字の頻度と異なり語数

図19，図20はx軸に文字の頻度，y軸に文字の用いられる異なり語数をとり，各文字についてプロットしたものである。図19は漢字のみを表したもので，図中の“◇”印は同一頻度，同一異なり語数のために重複する点の数を示している。“◇”印の対角線の長さは，グラフの一目盛り当り500個の点に相当する。

文字の異なり語数は文字の出現頻度よりも多くなることがないので，すべての点は直線$y=x$よりも下の部分に存在する。直線$y=x$上の点は，その文字のすべての出現において，異なった語に使われていたことを意味する。右上方には出現頻度，異なり語数ともに大きな文字が分布する。直線$y=1$上の点は，単独の文字として，あるいは単一の語の中でのみ用いられた文字である。

表3に異なり語数が少なく使い方が限られている漢字（異なり語数が3語までのもの）を出現頻度順に20位まで示す。一方，一つの漢字が出現するごとに異なった使い方がされるような，出現頻度に対して異なり語数の大きい漢字を表4に示す。これは直線$y=x$と$y=0.5x$の間に存在する漢字，すなわち，2回に1回以上は異なった語に用いられるような漢字を頻度順に20位まで示している。表1～表4を参照することにより，図19に示された点がどの漢字

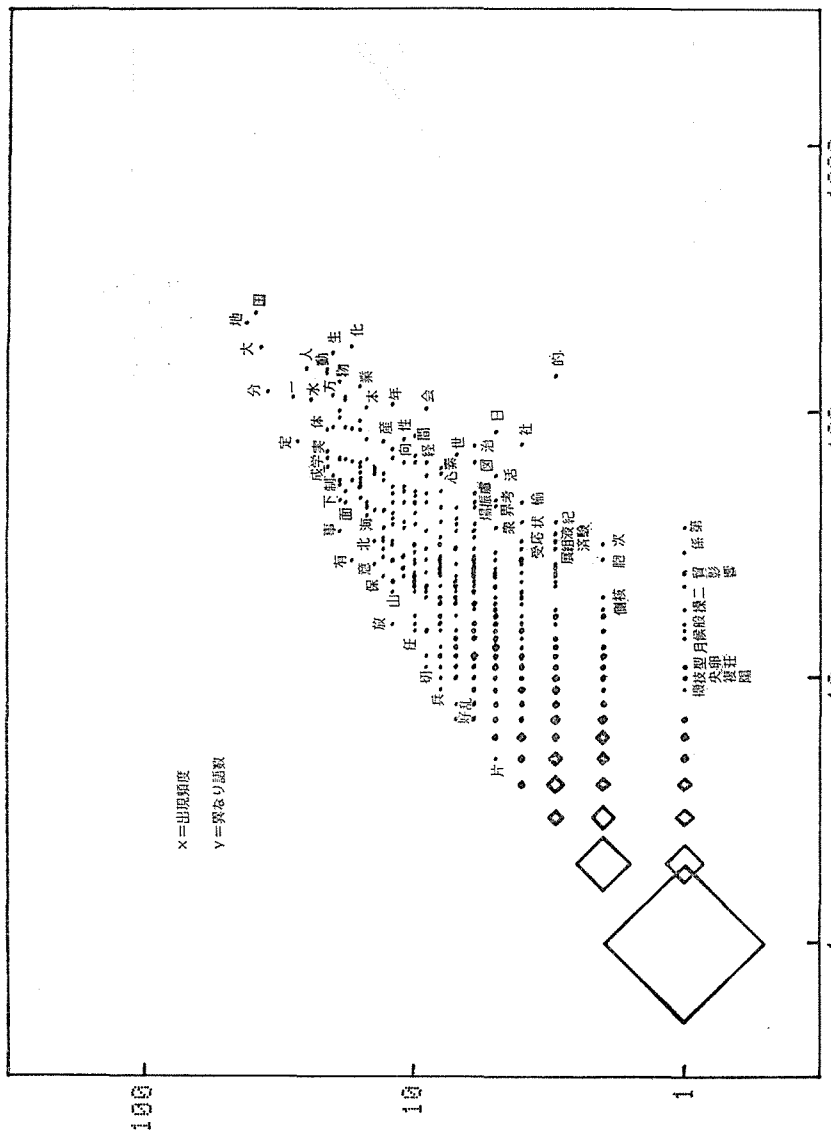

図19　漢字の出現頻度と異なり語数

に対応するか調べることができる。

表3　異なり語数の少い漢字

	異なり語数＝1			異なり語数＝2			異なり語数＝3		
	文字	用語	頻度	文字	用　語	頻度	文字	用　語	頻度
1	第	第	37	次	次，次い	32	的	的，目的，公的	138
2	係	関係	30	胞	胞胚，細胞	28	紀	紀元，世紀，年紀	39
3	影	影響	25	核	核，無核	20	液	液，溶液，だ液	35
4	響	影響	25	側	側，側面	18	験	実験，試験，経験	35
5	貿	貿易	25	必	必要，必ず	17	組	組，組織，組み	33
6	二	二	22	昭	昭和，斉昭	17	済	経済，救済，決済	33
7	操	操作	18	織	織，組織	17	展	展開，発展，進展	31
8	般	一般	16	企	企業，企て	15	胚	胚，胚柄，胞胚	27
9	候	気候	15	試	試験，供試	15	態	態度，状態，事態	26
10	月	月	14	幕	幕布，幕末	15	程	程度，過程，方程	26
11	型	型	11	可	可能，不可決	13	由	自由，理由，経由	25
12	荘	荘園	11	報	報国，情報	13	達	達する，発達，伝達	24
13	卵	卵	11	刺	刺激，風刺	13	憲	憲法，憲章，立憲	23
14	技	技術	10	件	条件，事件	12	洋	洋，海洋，東洋	22
15	央	中央	10	何	何，幾何	12	再	再生，再建，再開	18
16	複	複雑	10	帝	帝国，皇帝	12	川	川，河川，徳川	17
17	陽	太陽	10	雑	雑貨，複雑	12	働	働く，労働，作働	17
18	械	機械	9	荷	荷，電荷	12	値	値，価値，数値	17
19	誠	誠	7	蒸	蒸気，蒸発	12	与	与え，与える，貸与	17
20	鮮	朝鮮	7	非	非，非常	11	濃	濃度，濃霧，濃く	16

表4　相対的に異なり語数の多い漢字

	文字	頻度①	異なり語数②	②／①		文字	頻度①	異なり語数②	②／①
1	事	36	19	.53	11	田	16	8	.50
2	有	28	17	.61	12	交	16	8	.50
3	意	27	14	.52	13	任	15	10	.67
4	保	24	13	.54	14	調	15	9	.60
5	通	23	12	.52	15	書	15	9	.60
6	流	23	12	.52	16	路	15	8	.53
7	山	21	12	.57	17	設	14	7	.50
8	安	17	10	.59	18	固	14	7	.50
9	放	16	12	.75	19	半	14	7	.50
10	無	16	10	.63	20	村	13	8	.62

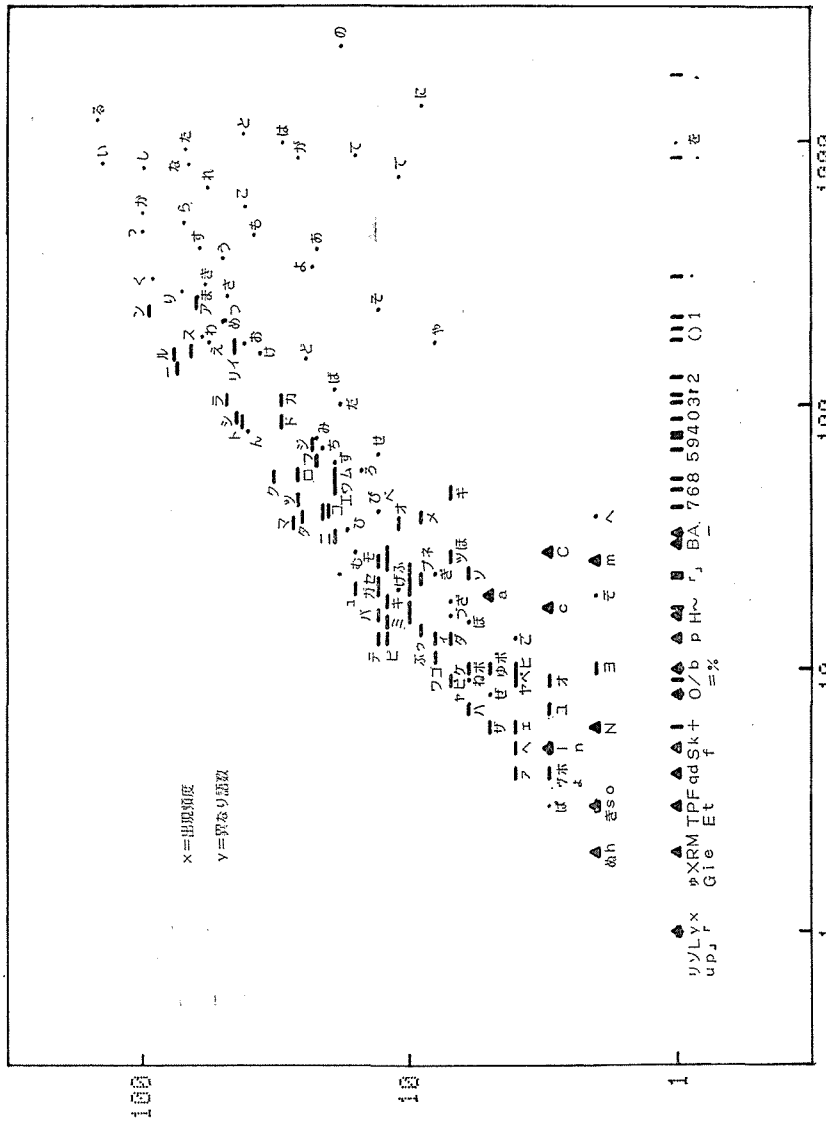

図20　平仮名，片仮名，英字，数字，記号の出現頻度と異なり語数

図20は漢字以外の文字（平仮名，片仮名，英字，洋数字，記号類）に関して頻度と異なり語数の関係を表したものである。このグラフは重複する点の数が比較的少ないので，字種別に異なった記号を用いている。図中の点は平仮名を示し，短い横線は片仮名，三角形は英文字，短い縦線は洋数字および記号類を示している。

平仮名は使用頻度，異なり語数ともに高い右上方を中心として左下方へかけて広く分布している。特異なものとして「を」は，格助詞以外の用法がないので，異なり語数が1で使用頻度の高い点となって現われている。

片仮名は外国の人名，地名など固有名詞に用いられることが多いので，使用頻度の割りには異なり語数の多い範囲（直線 $y=x$ と直線 $y<0.3x$ の間）にほとんどのものが分布する。

英字は記号的に使われることも多く，使用頻度，異なり語数ともに少ない右下方に分布する。洋数字および記号類は1字で1M単位をなすので，異なり語数が1の線上に並ぶ。

5. 文字列における語の境界

(1) 語における文字の位置

分かち書きの習慣のない日本語文を計算機で処理する場合に，文字列の中に語の境界を見出し，語を単位として取り出すことが必要になる。文字が語の中で使用されるとき，特定の文字は語の先頭にだけ，あるいは末尾にだけ用いられるといった性質があれば，語の境界を見出すのに有効になる。

図21は異なり語数（M単位語）を6個以上持つ文字，すなわち6通り以上の語としての使われ方のある文字（399個）について，その文字がM単位語のどのような位置にくるかを調べたものである。x 軸はある特定の文字が，その文字の出現したすべての語において，先頭の文字になっていた割合を示し，y 軸はその文字が同じく末尾の文字になっていた割合を示している。図中の記号で点は漢字を，短い横線は平仮名を三角形は片仮名を示している。

このグラフにおいて点で示された漢字は座標（100, 0）と座標（0, 100）を結ぶ線よりも上方に存在する。これは漢字を示す点の x 座標と y 座標を加えた

ものが100以上であること，すなわち先頭の文字になる可能性と末尾の文字になる可能性を加えたものが100%以上であることを意味する。これは漢字を含むM単位語の多くが2文字以下で構成され，語の先頭の文字か末尾の文字かその両方になっていたこと，また2文字以上のM単位語は先頭か末尾が漢字となり，中ほどの文字は漢字になりにくかったことを予想させる。特に座標（100, 0）と座標（0, 100）を結ぶ線上に存在する漢字の多くは2字漢語で使用されたものと思われる。

平仮名は語の先頭の文字になりやすいものから末尾の文字になりやすいものまで広く分布している。右上方に分布する一群の平仮名（「に」，「は」，「や」，「て」，「が」，「で」，「の」）は助詞として使われ，単独でM単位をなすことの多い文字である。

片仮名は左下方に集まっている。これは片仮名により構成されるM単位語は比較的文字数が多く，先頭の文字や末尾の文字になる割合が相対的に減少し，語の中ほどの文字となる割合が増大したことに起因する。

図 21 の周辺には，すべての出現において語の先頭になった文字(右端)，一度も先頭にならなかった文字(左端)。すべて語の末尾となった文字(上端)，一度も末尾とならなかった文字(下端)を示している。

語の先頭にはなりやすいが末尾にはなりにくい文字として，図 21 のグラフにおいて $y \leqq x-80$ の範囲に存在する文字を異なり語数の多い順に示す。

お，よ，あ，ひ，同，そ，不，変，ふ，多，特，最，無，知，各，
増，少，ほ，比，低，観，広，基，武，急，天，異，ハ

一方，語の末尾にはなりやすく先頭にはなりにくい文字として $y \geqq x+80$ の範囲に存在する文字を異なり語数の多い順に示す。

る，っ，り，ん，ム，回，業，ろ，び，族，素，和，路，命，身，
料，治，域，想，害

図 22 は図 21 と同じ文字（6通り以上のM単位語に使われる）を対象として，W単位語[3]に関して同じ調査を行ったものである。助詞としての用法の多い少数の平仮名を除いて，大多数の文字は右下方へと移動する。これはW単位語がM単位の結合された形の比較的長い文字列からなるためで，各文字が語の

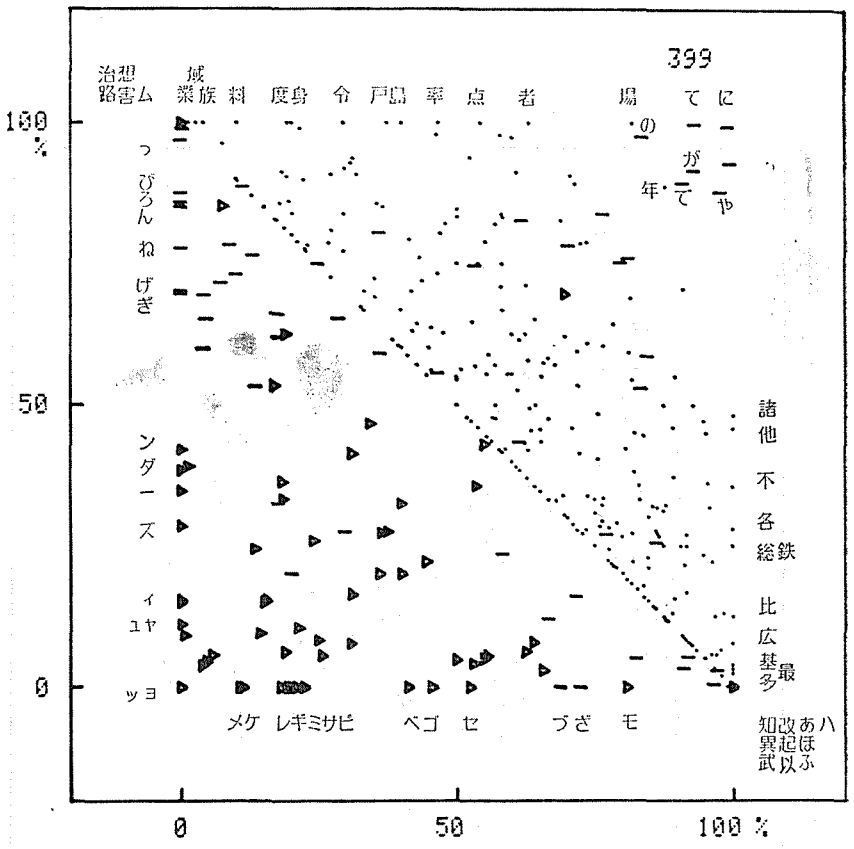

図21　M単位語における文字の位置

先頭や末尾にくる割合が相対的に減少したことによる。

図22の周辺には，すべての出現において語の先頭となった文字(右端)，一度も先頭にならなかった文字(左端)，すべて語の末尾となった文字(上端)，一度も末尾とならなかった文字(下端)を示している。

W単位語の先頭にはなりやすいが末尾にはなりにくい文字として，図22において $y \leqq x-80$ の範囲に存在する文字を異なり語数の多い順に示す。

お，よ，あ，ひ，同，自，反，そ，公，モ，変，ふ，比，広，振，
以，基，思，改，武，天，思，起，ハ

一方，W単位語の先頭になりにくく末尾になりやすい文字として $y \geqq x+80$ の範囲に存在する文字を異なり語数の多い順に示す。

る，う，て，路，者，命，率，料，域，島，害

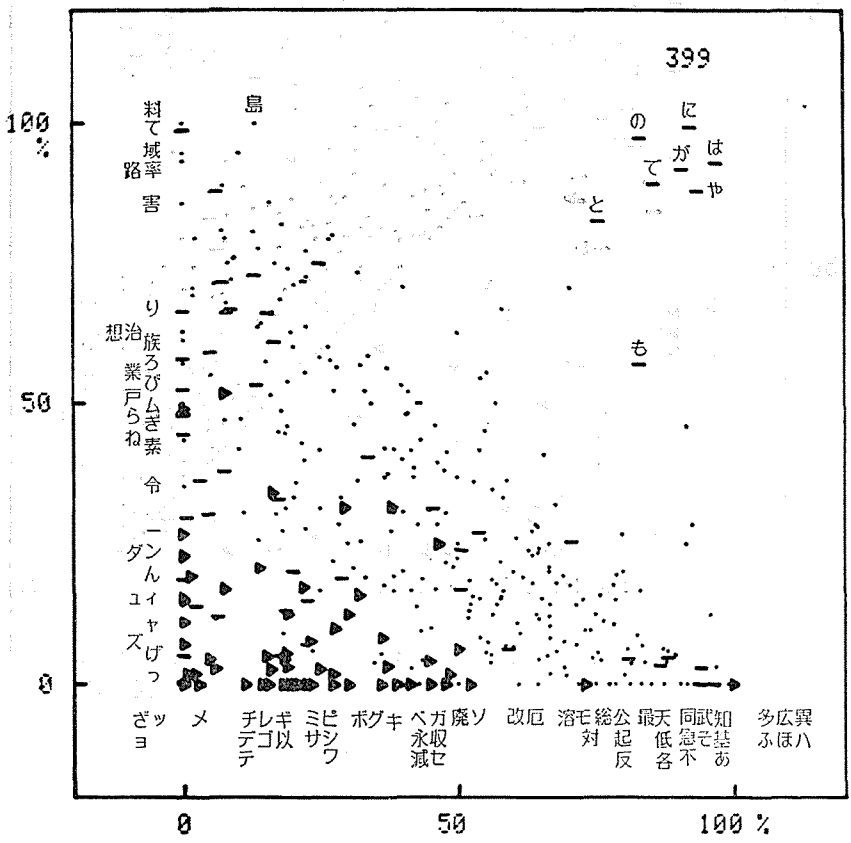

図22　W単位語における文字の位置

(2)　字種の境界と単位

文節は一般的には漢字に始まり仮名に終ると考えられる。これは体言を含む文節が漢字で書かれる名詞に始まり平仮名で書かれる助詞に終ること，また用言を含む文節が漢字で書かれる動詞や形容詞に始まり平仮名で書かれる活用語尾や助詞，助動詞などを従えることに基づいている。ここでは文節よりも短いM単位，W単位に関して字種の境界と語の単位との関係について調べる。

(i) M単位の場合

図23は全漢字データの中で，少なくとも1回は平仮名に続くことがあり，しかも，直前が少なくとも1回はM単位語の境界（W単位語の境界を含む，以下同じ）になる漢字902個（延べ15461個）について，字種の変化と語の境界の関係を調べたものである。x軸は漢字の直前に平仮名が来たときに，その漢字と平仮名との間がM単位の境界となっていた割合を示している。y軸は同じ漢字の直前がM単位の境界となるときに，直前の文字が平仮名であった割合を示している。図中の“◇”印は各々の漢字を示す。対角線の長さは漢字の出現頻度を表し，グラフの目盛で10%の長さが出現頻度100回に相当する。棒グラフはx軸，y軸ともに，10%ごとの区間に分け，各区間に存在する文字の出現頻度を加え合せたもので，グラフ全体の延べ文字数に占る割合を表している。

各区間の端点は端点よりも小さい方の区間に属させている。ただし，0%にかぎり0～10%の区間に入れている。

x軸に関する棒グラフからは，平仮名が漢字に続くとき，ほとんどの場合がM単位の境界になっていることがわかる（棒グラフのxが90%以上のときyは97.9%，xが100%となるときyは95.8%）。境界とならなかったごく少数の漢字についてKLIC（文字KWIC）を用いて原因を調べると，次のような仮名漢字まぜ書き語に使われていたことがわかる。

大：ぼう大	流：かん流	岩：でい岩	類：そう類
水：こう水	条：か条	積：たい積	配：こう配
開：へき開	動：はく動	星：わい星	液：だ液

y軸に関する棒グラフからは，M単位の境界だからといって，漢字の前に平仮名が来るとは限らないことがわかる。漢字にはM単位語の先頭の文字となるときに，平仮名に続きやすいものから続きにくいものまで広く分布している。

表5には座標（100，100）に存在する漢字で，6通り以上の異なった語に用いられるものを示す。これらの漢字は平仮名に続くとき，漢字の直前が100%M単位の境界となり，また直前がM単位の境界となるときは100%平仮名に続くようなものである（たとえば表5の欄(A)に示される語は漢字で書かれた語に続いて複合語を構成することがなかった。言いかえると，M単位の境界にな

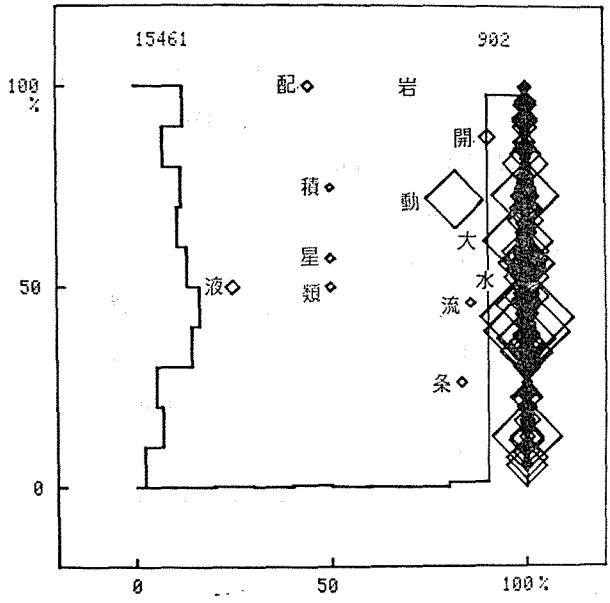

図23　仮名に続く漢字とM単位の境界

らないときは直前の文字が平仮名とならず，直前の文字が平仮名でなければM単位の境界とならないものである。（表5の欄(B)参照）

表6には図 23 に現れなかった漢字，すなわち一度も平仮名に続いたことのない漢字（欄(1)）と一度もM単位語の先頭にならなかった漢字（欄(2)）が用いられた語を示す。（異なり語数が6以上のもの）

表5　座標（100, 100）に存在する漢字

文字	頻度	(A)	(B)
異	11	異形，異化，異常，異なり，異なる，異なっ	——
起	9	起源，起き，起きる，起こ，起こる，起こし	——
少	17	少数，少量，少ない，少なかれ，少なく，少し	多少，減少
急	12	急(傾斜)，急進，急速，急激，急減	緊急
設	14	設置，設定，設備，設け	建設，敷設，施設
導	13	導体，導入，導か，導く	指導，主導，誘導
好	7	好(／都合)，好転，好ましい，好み	愛好，友好，(吉田／)兼好
断	10	断固，断層，断裂，断ち	横断，切断，判断，不断
属	29	属さ，属し，属する	金属，従属，隷属
求	26	求人，求め，求める	追求，要求，欲求

廃	10	廃絶，廃虚，廃藩	荒廃，全廃，撤廃
離	11	離反，離れ，離す	電離，分離，遊離，距離
乱	8	乱，乱後，乱れ	戦乱，争乱，反乱，混乱
任	15	任命，任免，任ぜ	信任，補任，遥任，一任 責任，勅任，赴任
命	27	命，命じ	生命，革命，天命，任命 立命
兵	9	兵部，兵権	親兵，徴兵，傭兵，府兵 皆兵，騎兵
来	22	来日，来航	以来，本来，未来，将来 旧来，渡来，従来
身	10	身	前身，自身，修身，出身 一身，単身

表 6　図23に現れない漢字

	文字	頻度	用いられた語
(1)	業	126	企業，工業，農業，作業，産業，分業，事業，失業，就業，綿業 商業，卒業，林業，農牧業，(鉄鋼／)業
	族	49	華族，家族，部族，士族，民族，種族，貴族，皇族，一族，(一／)族
	域	38	海域，全域，地域，流域，領域，(地中／海／)域
	令	17	律令，県令，司令，指令，禁令，勅令，(分割／)令
	料	10	衣料，原料，燃料，肥料，食料，香料，(小／作／)料
	戸	8	門戸，水戸，江戸，平戸，木戸，(数／)戸
(2)	治	75	政治，明治，内治，統治，官治，自治
(2)	素	62	水素，塩素，元素，炭素，酸素，要素，色素，窒素
(2)	想	27	空想，構想，思想，予想，理想，観想
(2)	路	15	経路，水路，回路，海路，道路，一路，販路，航路
(2)	害	7	公害，無害，被害，災害，弊害，侵害

図 24 は直後に平仮名が続き，しかも直後がM単位の境界（W単位を含む）となることのある漢字 804 個（延べ 15490 個）について，字種の変化と語の境界の関係について調べたもので，x 軸は漢字の直後に平仮名が来たとき，その漢字と平仮名との間がM単位語の境界になる割合を示し，y 軸は漢字の直後がM単位の境界になる場合に直後の文字が平仮名になる割合を示している。

捧グラフは図 23 と同じように，10% ごとの区間に入る文字の出現頻度を加え合せたもので，グラフ全体の延べ文字数（15490 個）に占る割合を表している。図 23 ほど極端ではないが，後に仮名が来ればM単位語の境界になること

が多いことがわかる（棒グラフはxが90%以上のときyは69.6%，xが100%のときyは64.0%）。

表7は座標（100，100）に存在し，異なり語数が6以上の漢字を示す。これは後に仮名が来れば，必ず直後がM単位の境界となり，直後がM単位の境界となるときは，必ず後に仮名が続くような漢字であり，したがってこの調査では後に漢字で書かれる語を従えて複合語を構成することのないような漢字である。（欄（B）参照）

これはまた直後がM単位の境界とならないときは，後に仮名が続くことがなく，後に仮名が続かなければM単位の境界とならないような漢字と言い換えることができる。（欄（A）参照）

直後がM単位の境界となるとき，後に仮名が続く漢字をさらに調べるため，直線 $y=100$ 上の点を求めると次のようになる（$x\neq100$，xの大きい順）。

現，流，自，適，富，細，熱，納，加，開，調，書，求，変，採，思，

確，増，比，減，失，少

これらは，いずれも用言の語幹としての用法を持つため，平仮名に続くとき必ずしもM単位とならず，$x\neq100$ となっている。これらの文字に限らず，図24においてxが100%となっていないものの多くは，用言としての用法を持つものと思われる。

表7　図24の座標（100，100）に存在する漢字

文字	頻度	(A)	(B)
特	32	特徴，特定，特産，特質，特有，特色，特殊，特権	特(／に)，独特
基	26	基盤，基部，基本，基礎，基準	基(／づく)
他	24	他人，他国，他方，他事，他者	他
温	24	温度，温帯，温泉，温暖	高温，常温
鉄	12	鉄鋼，鉄砲，鉄器，鉄道	鉄，銑鉄
鉱	19	鉱山，鉱産，鉱物，鉱石	鉱，鉄鉱，銅鉱
真	12	真空，真実，真剣	真，写真，(管原／)道真
廃	10	廃絶，廃虚，廃藩	荒廃，全廃，撤廃
論	6	論議，論争	論，世論，理論，推論
圧	10	圧迫，圧倒	圧，電圧，気圧，弾圧
害	7	——	公害，無害，被害，災害，弊害，侵害

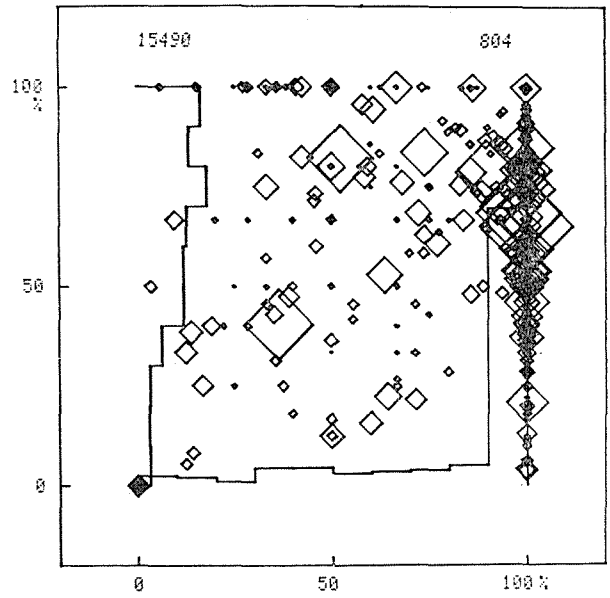

図24　仮名を従える漢字とM単位の境界

表8に座標（0，0）に存在する漢字で異なり語数が6個以上のもの，すなわち，6通り以上の異った使われ方をして，後に仮名が続くときにはけしてM単位とならず（表8欄(B)），M単位となるときはけして仮名が続くことがなかった（表8欄(C)）漢字を示す。

表8　図24の座標（0，0）に存在する漢字

文字	頻度	(A)	(B)	(C)
同	53	同意，同化，同居，同年，同国，同志，同時，同様，同心，同一，同量，同盟	同じ，同じく	共同(／体)，同(／振動)
多	42	多数，多額，多様，多少，多片	多かっ，多い，多く，多かれ	多(／細胞)
最	32	最短，最大，最近，最古，最後，最高，最初，最低	最も	最(／南端)
低	20	低調，低率，最低，低利	低く，低い	低(／緯度)
広	13	広大，広範，広島	広く，広い，広がる	広(／範囲)

表9は図24に現われなかった漢字，すなわち一度も直後に平仮名が続かなか

ったもの（欄(1)）と，一度も直後がM単位の境界とならなかったもの（欄(2)）の内で，6通り以上の異なった語に用いられたものを示す。

表 9　図24に現れなかった漢字

	文字	頻度	用いられた語
(1)	公	31	公家，公事，公共，公的，公転，公布，公社債，公職，公害，公債 公共，公園，奉公(／人)
	不	28	不当，不足，不在，不平，不法，不満，不断，不便，不安，不正，不可決，不(／可能)
	各	32	各人，各地，各部，各省，各片，各国，各(／方面)
	永	12	永遠，永続，永久，永代，永世，嘉永(／1)，寛永(／2)，貞永(／式目)
	急	12	急進，急速，急激，急減，緊急(／勅令)，急(／傾斜)
	天	12	天子，天文，天平，天命，天皇，則天(／武后)
	総	8	総合，総力，総覧，総称，総額，総(／人口)
(2)	以	36	以後，以来，以上，以下，以外，以前
	武	16	武家，武士，武力，武器，武后，武蔵
	改	18	改新，改正，改革，改善，改修，改め
	起	9	起源，起め，起きる，起こっ，起こる，起こし
	知	21	知性，知事，知恵，知識，知っ，知ら，知る，知ろ
	異	11	異形，異化，異常，異なる，異なっ，異なり

(ii)　W単位の場合

図 25 は図 23 に対応するもので，少なくとも 1 回は平仮名に続くことがあり，また，少なくとも 1 回は直前が W 単位語の境界となる漢字 896 個（延べ 15401個)について，字種の変化と語の境界の関係を調べている。x 軸は漢字の直前に平仮名が来たときに，その漢字と平仮名との間がW単位の境界となっていた割合を示す。y 軸は同じ漢字の直前がW単位の境界となるとき，直前の文字が平仮名であった割合を示している。

x 軸に着目すると，平仮名に続く漢字はほぼW単位語の先頭であると見なすことができる（棒グラフの x が 90% 以上の部分の y は 92.0%，x が 100% となるときの y は 83.5%)。M単位の場合の方が多いのは，W単位の境界はM単位の境界でもあることによる。x が 100% とならなかった漢字について KLIC を用いて原因を調べると次のようなW単位語に用いられていたことがわかる。

x の小さい方から順に 90% までの区間に存在する漢字について用語の一部

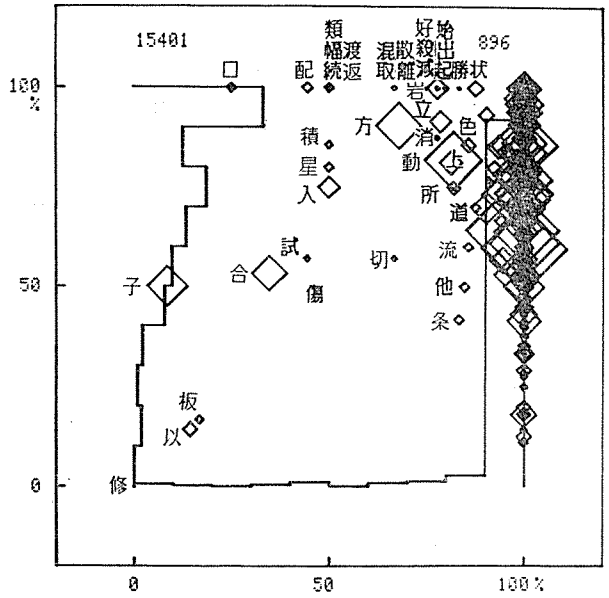

図25　仮名に続く漢字とW単位の境界

を示す。斜線はM単位の境界を示す。

修：合わ／せ／修め／なけれ／ば　子：振り／子

以：それ／以下，それ／以後，…　板：くし／板，条こん／板

口：割れ／口，きり／口　合：重なり／合う，組み／合わ／せ，

試：目／盛り／つき／試験／管　　打ち／消し／合っ／て，…

配：こう配，軸／こう配　傷：ひっ／かき／傷

入：ひき／入れ，取り／入れ　星：わい星

積：たい積　続：保ち／続け，生き／続け

返：くり／返さ／れる，…　渡：譲り／渡さ／れ／て

幅：振れ／幅　類：そう類

切：ぬぐい／切れ／て，抜け／切　取：受け／取っ／た，受け／取る

　れ／ない　離：切り／離す

混：振り／混ぜ／ながら，…　散：まき／散らし

方：伝わり／方，考え／方，…　岩：でい岩

好：より／好み／する　　殺：焼き／殺す
消：打ち／消さ／れ，…　　滅：すり／滅らす
起：引き／起こし，…　　出：取り／出し，乗り：出し／て，
立：成り／立つ，打ち／立て／ら　　始：し／始め／た
動：はく動　　上：投げ／上げ／た，…
条：４／か条　　勝：打ち／勝っ／て
他：その／他　　流：かん流／し
色：条こん／色　　道：わかれ／道
状：平たん／状，貝／がら／状

M単位の場合xが100％ とならなかった漢字はいずれも語を構成する際，直前の漢字が表外漢字となるために仮名書きされたことに起因していたが，W単位の場合は，上の用語例をながめると，複合用言の後半の語に使われたもの，用言により連体修飾を受けた名詞に使われたもの，代名詞により修飾を受けたものなどによることがわかる。

y軸に着目すると，漢字の直前がW単位の境界となるときは直前の文字が平仮名である割合が大きくなっていることがわかる。M単位の場合は複合語を構成する要素として働くため，直前に漢字の来る割合が大きかったが，W単位の場合は直前に漢字が来る割合が減少するはずで，もっと平仮名の来る割合が増大しても良さそうな気がする。そこで，yが100％ とならない漢字は直前がW単位の境界となるときに，平仮名以外のどのような文字が来ていたのか，ＫＬＩＣを用いて幾つか調べて見ると，次のように句読点が来ている場合，中黒が来ている場合，英字，記号が来ている場合，漢字が来ている場合などであることがわかる。

板：まず，板Ｂから…　　他：…を求めて，他の
上：…られた。上のように…　　切：…である。切断面が…
子：親分・子分の…　　合：石油化学・合成繊維など…
以：20°Ｃ以上）…　　条：…ない刺激（条件刺激）を
方：鴨長明の「方丈記」の…　　色：それ以上色が濃くなら…

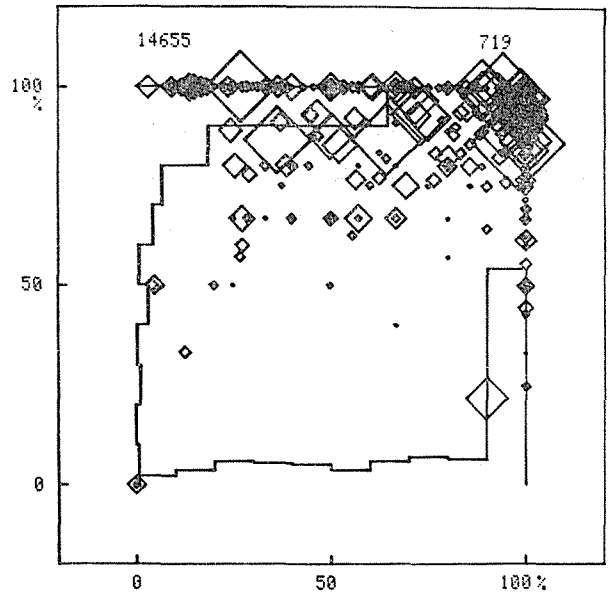

図26　仮名を従える漢字とW単位の境界

図26は図24に対応するもので，平仮名を従えることがあり，しかも直後がW単位語の境界となることのある漢字719字（延べ14655字）について調べたものである。x軸は漢字の直後に平仮名が来るとき，その漢字と平仮名との間がW単位語の境界になる割合を示し，y軸は漢字の直後がW単位となるときに直後の文字が平仮名になる割合を示している。

x軸に着目すると，漢字が平仮名を従えるときは直後がW単位の境界となる割合は図24のM単位の場合より減少していることがわかる。この原因を調べるため，xが30％以下となる漢字について，平仮名に続くときはM単位の境界となるがW単位の境界にはならないような用語例をKLICを用いて調べると次のようになる。サ変動詞の語幹がM単位では1語となり，W単位では仮名の語尾が続いて1語になる形が最も多いことがわかる。

大：大／はば　　重：重／さ　　現：現／に
高：高／める　　特：特／に　　見：見／て，見／やすく

視：岩倉／具視／ら　結：団結／し／て　用：作用／する
集：募集／し／た　開：展開／し／て　行：発行／する
意：用意／し　表：代表／する　加：参加／する
近：近／づい／た　通：開通／し／て　解：分解／し／て
使：行使／する　調：強調／し　増：激増／し
建：再建／さ／れ　進：促進／し　成：形成／さ／れ／た
少：減少／し　決：解決／する　半：折半／し／て
離：分離／し／た　振：共振／し／た　求：要求／する
割：分割／する

y軸に着目すると，図24と異なり，yが90%以上の部分の棒グラフがかなり大きくなっていることがわかる。これはW単位が文節から助詞を除いたものが多いため，漢字の直後がW単位の境界となるとき，平仮名が後に続くケースが増えることによる。yが50%以下となる漢字について，直後がW単位の境界となるとき，どのような原因で平仮名が続かなかったかKLICを用いて調べると，次のような文字連続であったことがわかる。

年：近年実験的研究が…，翌年共和政を…，1948年春，1948年復活し
軍：将軍義満と将軍義政…，アメリカ軍・イギリス軍…
西：西・北辺の…　山：登山・温泉保養…
小：中小・零細の…　振：共振・共鳴…
綿：大豆・綿・木材…　学：鉄鋼・化学・機械…
集：その作品集「金槐和歌集」…　建：トルコの再建」と立憲…
市：定期市（草市）がたった…　乱：東学党の乱（甲午農民…
見：伝達物質の発見）神経…　独：共和国（東独）を承認…
近：最近，北アメリカ…

図25において，漢字の直前とW単位の関係を調べたときと同じように，直後がW単位となりしかも仮名が来てない場合は，漢字が来る場合，中黒が来る場合，鍵や括弧が来る場合，読点が来る場合などがあることがわかる。

6. おわりに

東大大型計算機センター内に個人で借りることのできるディスクファイルは550（KB）となっている。これをソースプログラム，オブジェクトプログラムおよびデータ用として使うのであるが，TSSを便利に使うにはソース用とオブジェクト用の領域をあまり減らすことができず，データ用に使える分は80欄カードにして2000枚程度となる。

所内の計算機がTSS化されておれば，大量の原データに対する分析を端末から自由自在に行うことができるのであるが，残念ながらバッチ処理中心のシステムのためプログラミングの能率が極めて悪く，多くの試行錯誤を必要とする分析は行うことができない。一方，東大計算センターのシステムが大幅にレベルアップされないかぎり，原データ（教科書1/20データで数10(MB)）をセンタ内に置くことが不可能なので，あらかじめ分析に予想される情報をできるだけ多く含むように圧縮した2次データ（文字，頻度，異なり語，単位と文字環境など）を作成し，これをセンター内ファイルとして蓄え，TSS端末からの分析の対象とした。

ここで用いたプログラムはグラフを画くルーチン，座標を画くルーチン，グラフの点を調べるルーチンなどのほか，数値や文字の ASCII コード変換を行うルーチンなど約60本である。これら，教科書調査の1/20サンプリングデータを用いて開発したプログラムは，そのまま本データに対しても適用できるものであり，本データを用いて分析を行えばより良い結果を得ることができるはずである。なお，これは文部省科学研究費「言語解析を応用した日本語文修正処理の効率化に関する研究」の一端をなしている。

参考文献

(1) 野村雅昭：漢字仮名まじり文の文字連続，電子計算機による国語研究Ⅳ（国研報告46），1972

(2) 田中章夫：漢字調査における統計的尺度の問題，電子計算機による国語研究Ⅷ（国研報告59），1976

(3) 鶴岡昭夫：高校教科書用語調査の言語単位について，電子計算機による国語研究X（本報告書），1980