文字の計量調査

田中卓史（国立国語研究所）

1．はじめに

国立国語研究所の言語計量研究部では高校教科書の用語用字調査を行っている。教科書の全文章は長短2通りの語（W単位，M単位）に分割され，種々の付加情報と共に磁気テープに蓄えられている。（60万M単位）

ここでは，新しく導入したグラフィック端末（Tektronix 4006 東大計算センタTSS）を用いて，教科書データ中の文字を定量的・統計的に分析する。着目した文字集団の量的な側面は，文字の数，文字の出現頻度，文字の用いられる異なり語の数，語における文字の位置，文字列と語の境界である。

なお，パイロット的な試みであるので，教科書全データの1/20サンプリングデータを分析の対象としている。

2．教科別・字種別の文字数

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1種類と数えて）から構成されている。（図1，図2）

図3，図4は文字数（延べ）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，Y軸（縦軸）は字種別の割合を示す。

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：盤外記号　8：ピリオド(句点)　9：コンマ(読点)

J：日本史
W：世界史
E：政経
M：倫社
A：地理
P：物理
B：生物
G：地学
C：化学

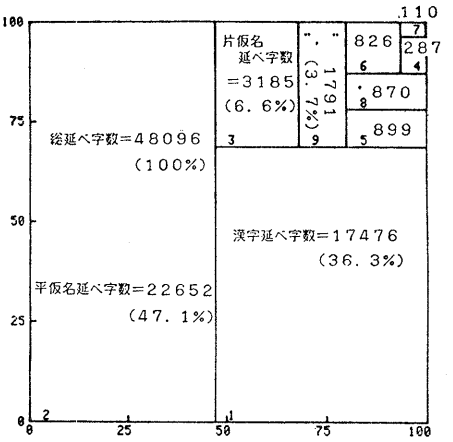

図1 延べ文字数の字種別割合

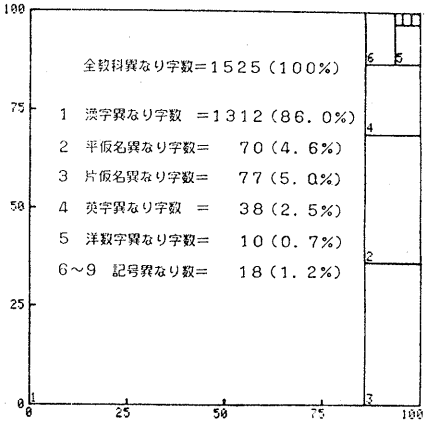

図2 異なり字数の字種別割合

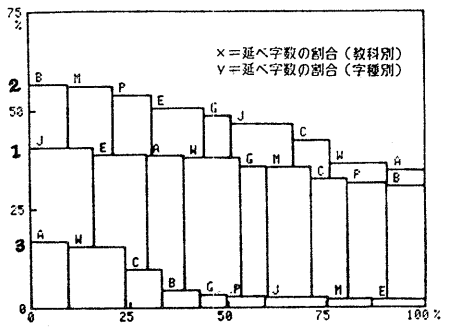

図3 漢字，平仮名，片仮名の教科別，字種別割合

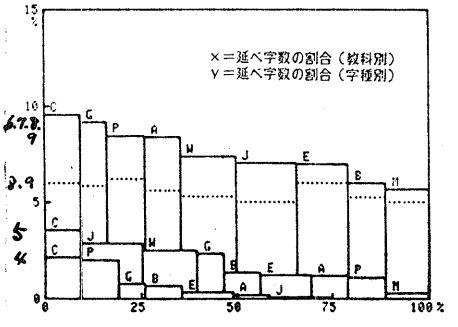

図4 英，数，記号の教科別，字種別割合

図5は異なり文字数の教科別・字種別割合を表したものである。各教科の幅(x軸)はそれぞれの異なり文字数に比例した大きさとなっている。x軸の100に相当する値は各教科異なり字数を単純に加えた値(5343字)である。一般に文字数の多い教科ほど漢字の占める割合が多くなっている。これは各教科の1/20程度の量のデータでは、漢字が十分飽和するまで出尽していないことに起因する。

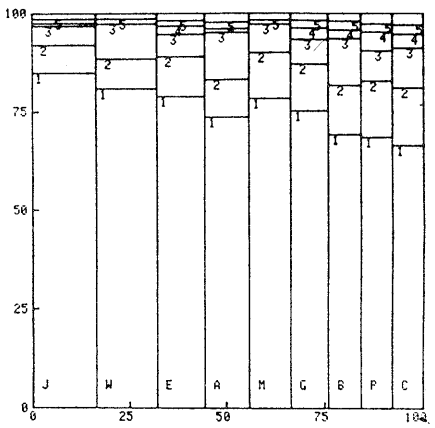

図5 異なり字数の教科別字種別割合

3. 文字の出現頻度

図6は各々の文字を出現頻度順に並べ、順位と頻度の関係を表したものである。出現頻度が10数回以下の部分は同一頻度の文字が多くなるので、グラフは横方向に延びる線分となっている。線分は左端の点が順位を、長さが頻度の等しい文字の数を表す。

図7は図6を構成する文字種の内訳けを見るため、x軸に全体順位、y軸に字種別順位を表している。最初に平仮名(2)が現われ、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、600位付近で平仮名の数を追い越す。これは片仮名が外国の人名・地名の表記に用いられたため、平仮名よりも多くの濁音、半濁音、よう音を表す文字が用いられたことによる。漢字は33位から現われ始め全体順位と共にほぼ傾きが1で増大する。これは5~60位以上の高順位では漢字が支配的となり、他の字種の文字がまばらに混じる状態となっていることを意味する。

図8は字種別に出現頻度と順位の関係を表したものである。平仮名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

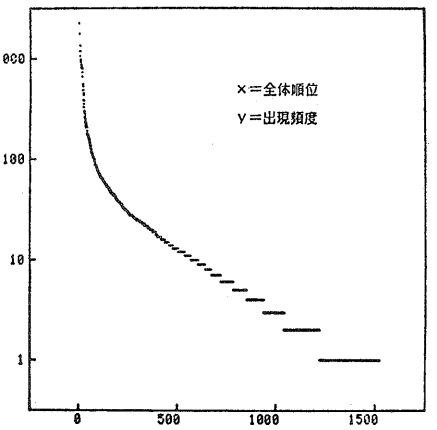

図6 文字の順位と出現頻度

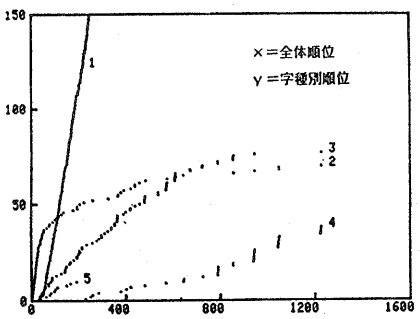

図7 文字の全体順位と字種別順位

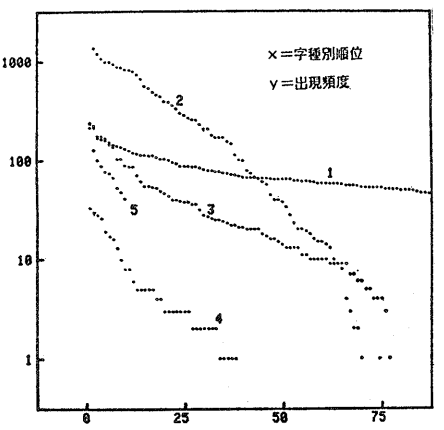

図8 字種別順位と出現頻度

回になるまで分布していることがわかる。

図9は文字の全体順位に対し，各字種ごとに累積延べ字数を求め，全延べ字数に占める割合を表したものである。平仮名(2)は極く若い順位までに全延べ字数の大部分を占める様子がわかる。各字種の最終順位における比率は図1に示したものに一致する。

図10は字種別順位と累積延べ字数の関係を表したものである。漢字は最終的に1312位において延べ17476文字となる。

図11は字種別順位と字種別累積比率の関係を示したものである。平仮名(2)，片仮名(3)共に，10位までで延べ文字数の50％を越えることがわかる。10位以上で平仮名の方が早く飽和の傾向を示すのは，若い順位の文字がより集中的に使われる傾向にあることを示す。

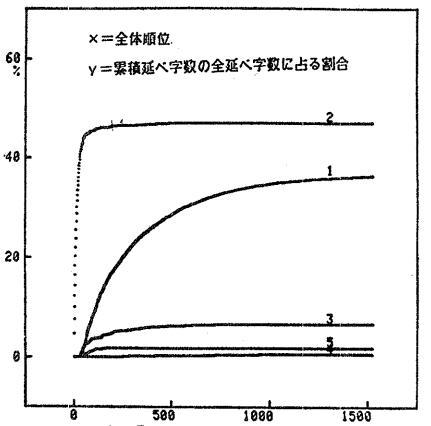

図9 全体順位と累積延べ字数の割合

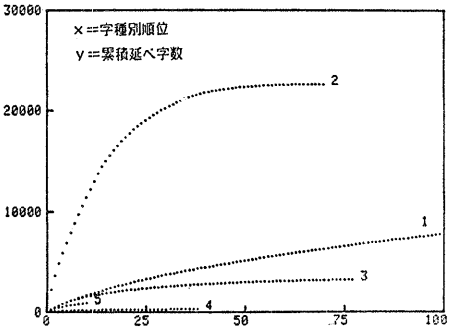

図10 字種別順位と累積延べ字数

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	e	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	143	53	¬	323	23
9	的	58	138	が	10	871	カ	72	105	b	570	10	6	163	48	ㄴ	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	t	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	—	—	—	ゥ	1221	1
18	年	71	108	っ	20	459	ウ	147	53	g	854	4	—	—	—	ゃ	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	—	—	—	—	—	—

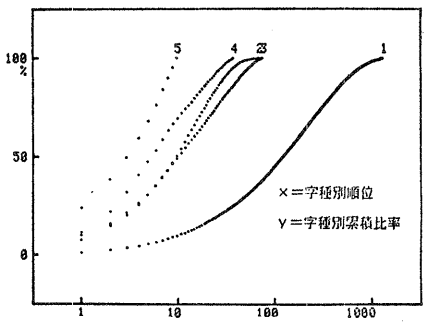

図11 字種別順位と字種別累積比率

4. 文字の用いられる異なり語数

一つの文字が幾つの異なった語の構成に用いられるかについて調べる。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」もまた各々の文字に関して定まる値である。図6～図11において，「頻度」を「異なり語数」で置換えると，同様なグラフ図12～18を得ることができる。

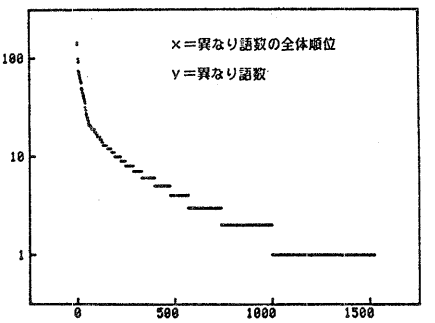

図12 文字の異なり語数と順位

図12は図6に対応するもので，文字の用いられる異なり語数とそれの順位の間の関係を表している。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る。次いで「い」，「か」，「っ」，「し」「ン」，「く」，「ル」，…と続く。

表2　文字の用いられる異なり語数とその順位

字種別 順位	漢字			平仮名			片仮名			英字		
		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	本	54	24	ら	11	70	リ	27	45	σ	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	16	21	S	1002	1

異なり語数の多い漢字「地」がどのような語に用いられているかKLIC（文字kwic）を用いて調べると，次の用例があることがわかる。括弧内は頻度数を示す。

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）

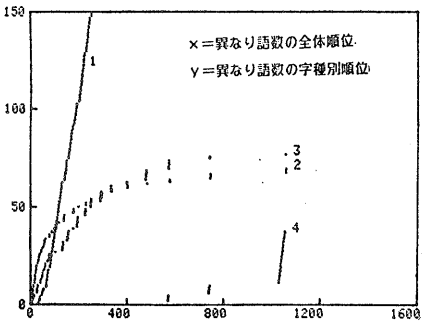

図13　異なり語数の全体順位と字種別順位

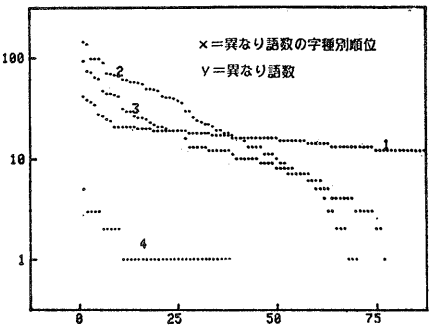

図14　異なり語数と字種別順位

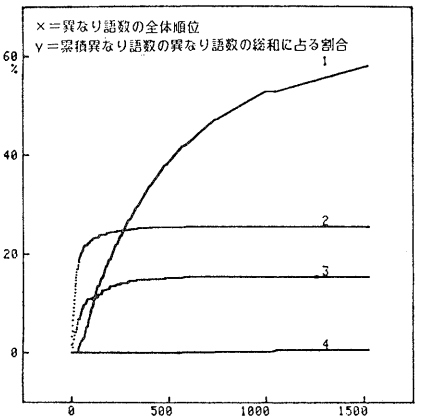

図15　全体順位と異なり語数の割合

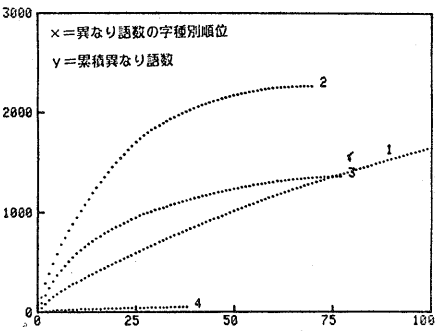

図16　字種別順位と累積異なり語数

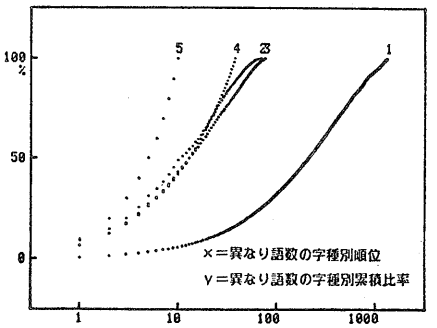

図17　字種別順位と累積比率

図18は漢字に関して，図19はその他の文字に関して各々の文字の出現頻度と異なり語数の関係を示したものである。図18の◇印は同一頻度，同一異なり語数のため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が並ぶ。KLIC を用いて，$y=1$ 上の漢字の用語を調べると次のようになる。

（頻度順に上位のものを示す）

第：第，係：関係，影，響：影響
貿：貿易，二：二，操：操作，般：一般
候：気候，月：月，型：型，荘：荘園
卵：卵，技：技術，央：中央
複：複雑，陽：太陽，械：機械
誠：誠，鮮：朝鮮

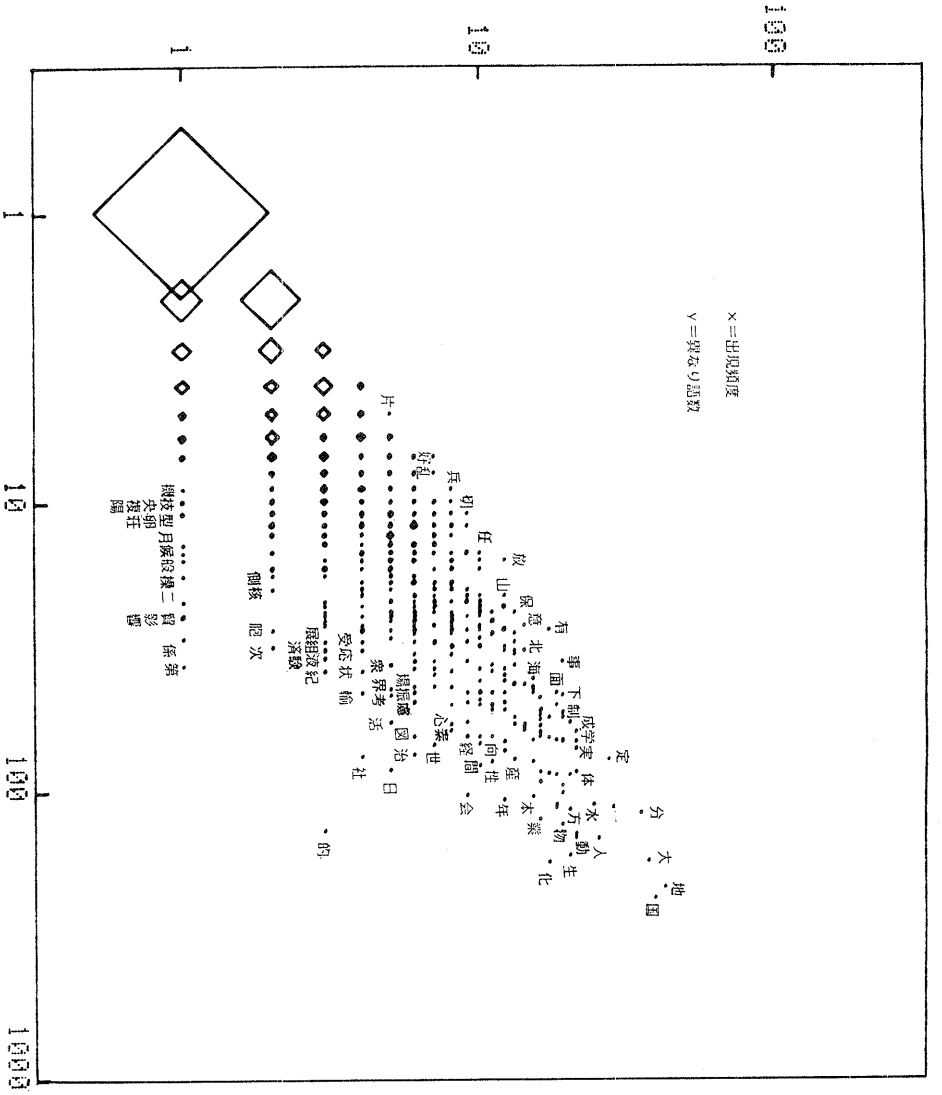

図18　漢字の頻度と異なり語数

図19は漢字以外の文字に関して頻度と異なり語数の関係を表したものである。このグラフは文字を表す点の重複が比較的少ないので，図18のように表わさず，字種ごとに異った記号を用いて表している。図中の点は平仮名，短い横線は片仮名，三角形は英字，短い縦線は洋数字および記号類を示す。平仮名は使用頻度，異なり語数ともに高い右上方を中心に分布する。特異な点として「を」は助詞以外の用法がないので異なり語数が1で出現頻度の高い点となって現れる。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，使用頻度の割に異なり語数の多い範囲に分布する。数字，記号類は1字で1M単位をなすので，直線 $y=1$ 上に並ぶ。

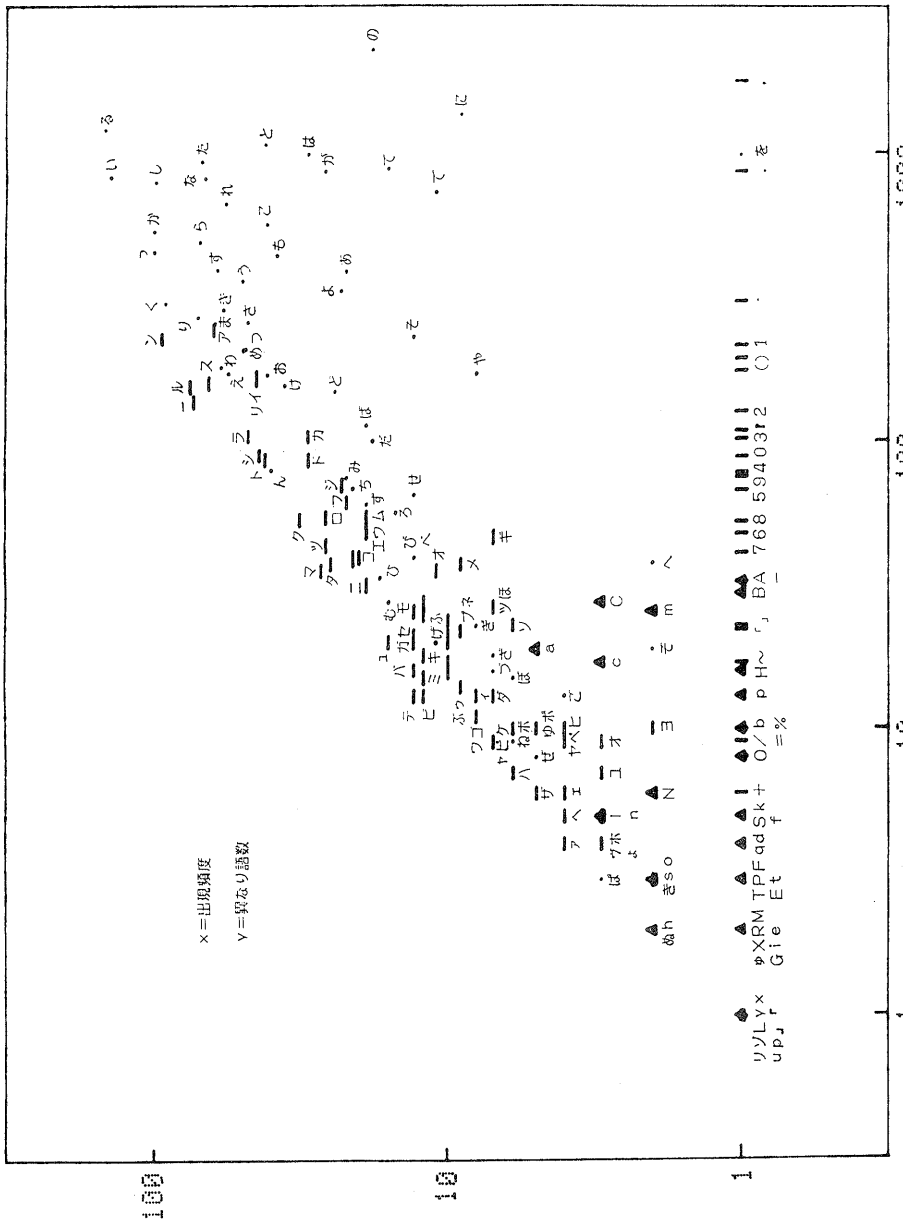

図19　平仮名，片仮名，英字，数字，記号の出現頻度と異なり語数

5．語における文字の位置

文字が語の中で使用されるとき，特定の文字は語の先頭だけ，あるいは末尾にだけ用いられるという性質があれば文字連続の中で語の境界を見出す際に有効になる。

図20は異なり語数（M単位）が6以上の文字，すなわち6通り以上の異なる語として使われる文字399個について，その文字が語のどのような位置にくるかを調べたものである。x軸は特定の文字が，その出現したすべての語（延べ）において先頭の文字になっていた割合を示し，y軸は同じく末尾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。点で示された漢字は，座標（100.0），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座標の値を加えたものが100以上であること，すなわち先頭になる可能性と末尾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。これを加えた値が100となる線上の漢字の多くは2字漢語でのみ使われたものと思われる。平仮名はグラフ全体に広く分布する。右上方に分布する一群の平仮名（「に」，「は」，「や」，「て」，「が」，「で」，「の」）は単独でM単位となる助詞として用いられる文字である。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語は比較的文字数が多く，語の先頭や末尾の文字になる割合が相対的に減少し，語の中程の文字となる割合が増大したことによる。

図の周辺には，文字の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。

図22は図21と同じ文字を対象としてW単位語に関して同じ調査を行ったものである。助詞としての用法を持つ平仮名（「に」，「は」，…）以外の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単位語の結合された比較的長い文字列からなるためで，相対的に語の先頭や末尾になる割合が減少したことによる。

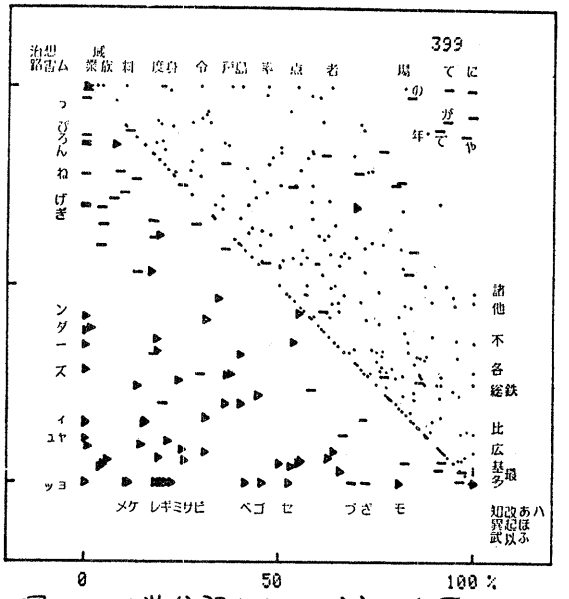

図20　M単位語における文字の位置

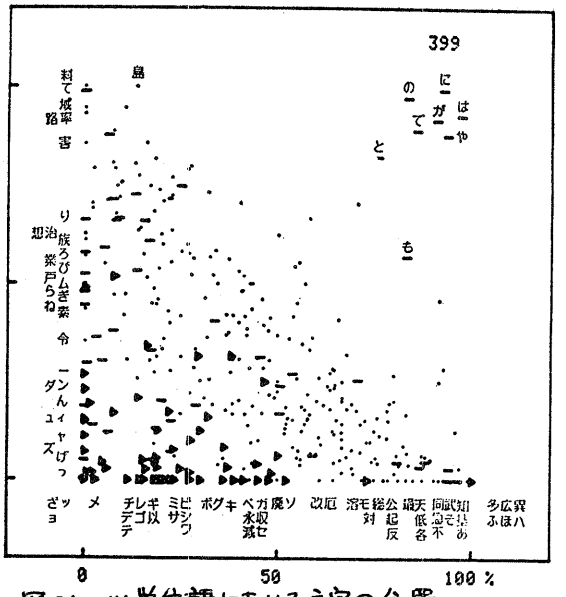

図21　W単位語における文字の位置

6. 字種の境界と単位

図22は全漢字データの中で少くとも1回は平仮名に続くことがあり，しかも直前が少くとも1回はM単位の境界（W単位の境界を含む）となる漢字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，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，y軸は同じ漢字の直前がM単位の境界になるときに，直前の文字が平仮名であった割合を示している。

図中の◇印は個々の漢字を示す。対角線の長さは漢字の出現頻度を表し，座標の10%の長さが100回に相当する。棒グラフはx軸y軸を10%ごとの区間に分け，各区間に存在する漢字の頻度を加え合せ，グラフ全体の延べ文字数に占める割合を表している。各区間の端点は端点より小さい方の区間に所属させている（0%に限り大きい方）。

x軸に関する棒グラフからは，漢字が平仮名に続くとき，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。境界とならなかった少数の漢字についてKLICを用いて原因を調べると，次のような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，動：はく動…

y軸に関する棒グラフからはM単位の境界だからと言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。

- 座標(100, 100)に存在し6通り以上の異なり語を持つ漢字：異，起，少，急，設，導，好，断，属，非，底，離，乱，住，命，兵，来，身
- 図22に現れない，即ち仮名に続かないかまたは直前がM単位とならない文字：業，族，域，令，料，戸，治，素，想，路，害，…

図23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）について調べたもので，x軸は漢字の直後に平仮名が来たときM単位の境界となる割合を示し，y軸は漢字の直後がM単位の境界となるとき直後の文字が平仮名になる割合を示している。

- 座標(100, 100)に存在し異なり語が6以上の漢字：時，基，他，温，鉄，鉱，真，底，論，正，宮
- 図23に現れない漢字で異なり語数が6個以上のもの：公，不，各，永，急，天，総，以，試，改，起，知，異

7. おわりに

文字に限らず，国研内に蓄積されている言語データに関し同様の分析を行う予定である。

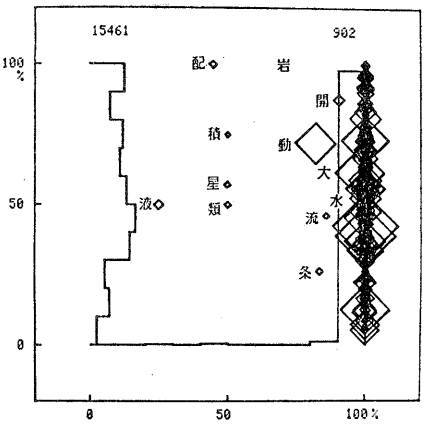

図22 仮名に続く漢字とM単位の境界

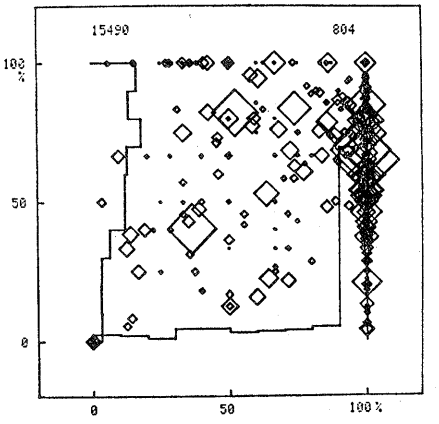

図23 仮名を従える漢字とM単位の境界