

出力形式の比較のために前回の形式も添付した。相違の詳細については4で述べる。なお、前頁の出力見本は正順ソートのものであるが、これと全く逆に第1キーの位置を右側にして逆順ソートした右順篇もある。述語をキーにして係り受け等を見る時によい。

○ 規模

品詞列集成の規模を数の上からみてみよう。

表1 品詞列集成の規模の比較

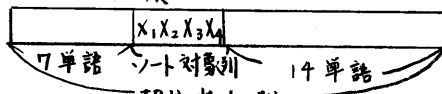
項 目	今回の内容	前回の内容
原文の語数	366,687語	984,977語
原文の文数	14,576文	3,0360文
印刷した行数(原本)	366,687行 ²⁾ ³⁾	なし
〃 (普及本)	73,337行 ⁴⁾ ⁵⁾	79,363行
3品詞列の種類	2,980種	5,078種
2品詞列の種類	327種	427種

○ 品詞列集成の使い方

利用者は、調べたい長さ(1~3まで)の品詞列で探す。その品詞列の中に特定品詞(助詞・助動詞・連体詞)が含まれていればその特定品詞にあたる単語の語形で探すこともできる。例えば、図1を〈名X助〉、〈名X助X助〉、〈名〉〈助の'〉〈形〉で探したりできる。

2. 形式の詳細

○ ソート順



—図3—

(ソートキー)

図1の①②③⑦ --- ソート対象列 $X_1 \sim X_4$ の品詞である(品詞キー)。

図1の④⑤⑥ --- $X_1 \sim X_3$ のいずれかが特定品

詞(助詞・助動詞・連体詞)の時のみの語形である(語形キー)。

ソートは上述の語形キーと品詞キーの組合せで行なわれる。キーの順序は①②③④⑤⑥⑦である。各キーのソート順は①②③⑦が

空名サナ付数ア連副接動用形陳助記読句并不

の品詞名順、④⑤⑥が国研コード順(ひらがな・カタカナのみのソート、これ以外は一括に扱う)である。

○ 単語の単位

国研の長単位に近い単位である。辞書テープの1長単位内に含まれる短単位列を後に述べるルールでまとめたものである。便宜的にこの単位からなる列と単語列とよぶ。

	また	金融	引き締め	に	批判的	な	多くの	銀行家	も、
(本書の単位)									
(長単位)									
(短単位…)									
前回の単位)									

図5のC

図5のA

図5の $B_1, B_2, \dots$

—図4—

。品詞分類

国研の品詞分類を基にキを加えたものである。

表2 品詞対応表

略品詞	品詞名	元の品詞		語例
		単品詞	多品詞	
空	空白品詞			
名	名詞	1, 2, 5, 8, 9	1と5, 1と8, 2と8, 1と9, 1とP	こと, もの, 東京都, 午前, …… 課税, 拒否, 汚職, …
サ	サ変動詞語幹	3		同じ, 明らか, 必要, …
ナ	形容動詞語幹	4	1とCとDとL	ごろ, 方, ら, 氏, …
付	付属語(接辞)	6		一, 一部, 一日, …
数	数詞	7		0, 1, 2, 7日, …
ア	数字(算用・ローマ)	X		この, その, 大きな, …
連	連体詞	D		いま, どう, 結局, …
副	副詞	C	1とC, 2とC, 9とC	では, また, しかし, …
接	接続詞・感動詞	A, B	AとC	いる, ある, し, …
動	動詞	E	DとE	考え, おり, 調べ, …
用	動詞類	+	1とE, 2とE	ない, よい, 多い, …
形	形容詞	L, -	LとP	た, だ, れ, …
陳	助動詞	P		の, を, ので, には, …
助	助詞	R		※MO, J, …
記	記号・符号	Y		,
読	読点(コンマ)	Y	読点, コンマ	.
句	句点(小数点)	Y	句点, ビリオド, 小数点	#
き	文区切り			>, ここで, the, …
不	無関係・不明	Z		

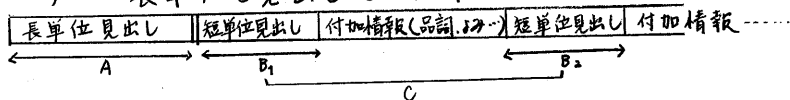
(品詞の説明)

多品詞…辞書テープ中の短単位に二つ以上の品詞がついているもののうち判別文法による品詞の区分けをしなかったもの。表示の略品詞外になる可能性があるのど、多品詞を含む品詞については注意が必要である。
2品詞一緒のもの…品詞分類の数を減らすために一緒にしたものである。多品詞同様、着目の単語が実際は何品詞なのか念頭におかなければならない。

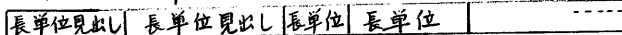
3. 作業手順

。原テープの形式

この作業は原テープの加工から始まるので原テープの形式を明示しておこう。
辞書テープ…長単位を見出しとし短単位ごとの付加情報がついたもの



原文テープ…長単位切りされた原文



—図5—

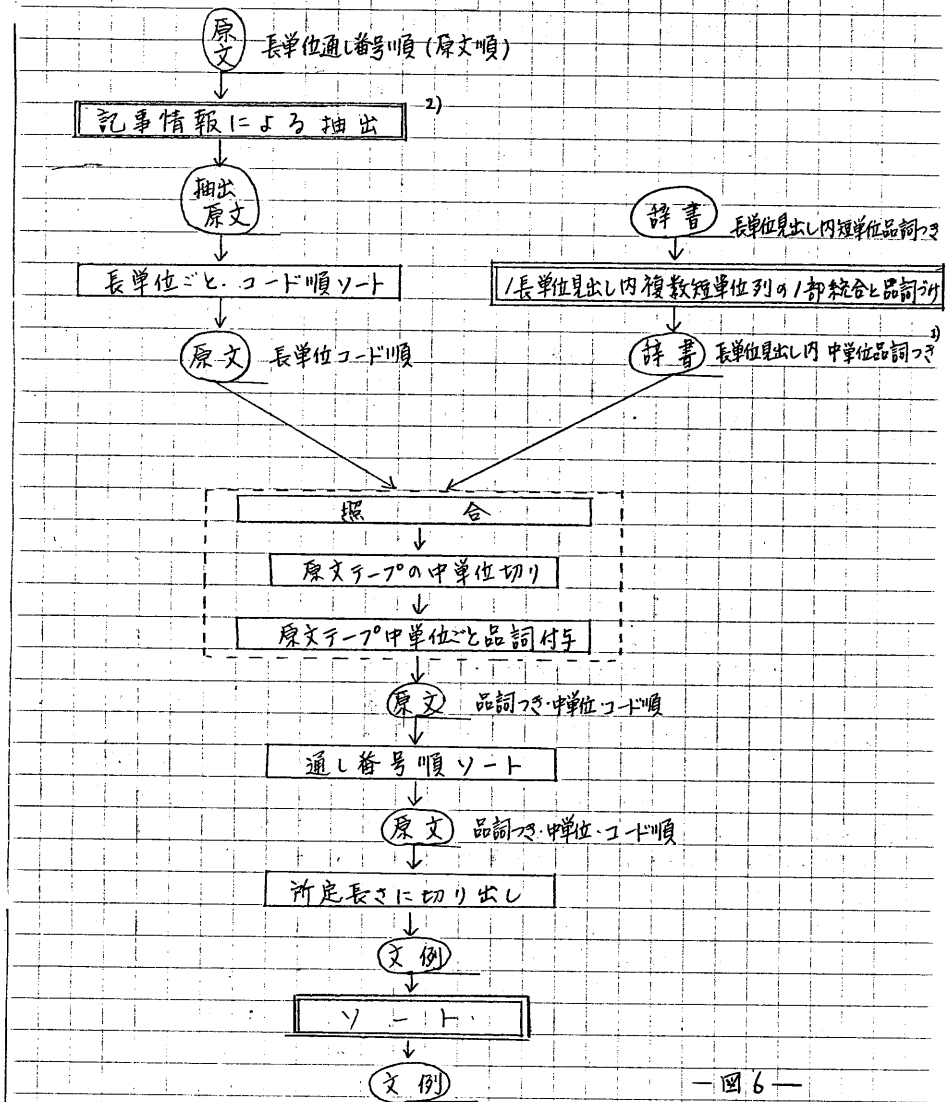
。作業概要

- (1) 上記の辞書テープを使って原文テープを辞書見出し右側の単位に切り直し、その単位ごとの品詞を付与する。
- (2) この品詞付き原文テープを「切り出し列」として所定の長さに切る。
- (3) この品詞付き「切り出し列」を所定の位置のキーでソートする。
- (4) ソートされた品詞付き「切り出し列」の集合を全部あるいはサンプリングして、所定の形式で出力する。

今日は、上記作業の前に以下の作業をし、原テープにキを入れた。

- ① 原辞書テープの書き換え——1長単位内(A)に含まれる短単位列(B1, B2)でまとめられるものは1単位(C)にする。
- ② 原文テープの記事情報によるデータの抽出——広告文等を除く。

以上の作業内容を図示する。下図の 部分は、前回作業への追加・変更部分である。



—図6—

4. 前回作業との相違点

この作業は前回の品詞列集成の改良版でもあるから、変更部分を明らかにしてみよう。

4.1 主な相違点

(1) 番組・広告記事等の削除 — 表6からもわかるように、前回のものは〈名詞の連続、〈数字〉の連続、〈名詞×記号〉の列等が上位を占め、構文的な資料としては不充分なところもあったので、対象データから番組・広告等を外した。

(2) 長単位内短単位列の書き換え — 複合語の構成要素等はまとめて、単語単位と長単位に近いものにした。〈名×名×助〉は〈名×助〉となり見やすくなった。表

1の品詞列異なりパターンの減少、表7の〈名詞×名詞〉、〈名詞×接辞〉等複合語構成要素列の減少となった。

(3) ソート順の変更——品詞だけでなく、助詞・助動詞等については単語の語形ソートを組合せにため、それらの用例や機能分析に有効である。

(4) 品詞づけの変更——前回、助詞は〈格〉、〈係・副〉、〈終・接〉の三つに分かれていたが、助詞については語形ソートを行なうので品詞の細分はやめて〈助〉とした。

(5) 出力形式の変更——キーになる単語列に少しでも長く前後文脈をつけるように工夫した。図1、図2によって出力形式を比較してみるとわかるように、

③ 切り出し通し番号の削除 ④ ソート対象列外の品詞名削除 ⑤ ソート対象列外の単語間空白削除
等が主な変更点である。

○ 表による形式の比較

以上の変更点を表の上でみると形式的には次のような相違となる。

表 3 品詞列集成の形式の比較

項 目	今 回 の 内 容	前 回 の 内 容
単 位	国研の長単位に近い単位	短単位
品詞分類 (空白品詞を含めない数)	19種	21種
ソートキー	3単語列の品詞とその3単語列内に含まれる特定品詞(助詞・助動詞・連体詞)を持つ語の語形	5単語列の品詞
切り出し列の形式		
1行の長さ	25単語	12単語
第1キーの位置	左(あるいは右)から 8単語目	左(あるいは右)から 3単語目
印刷文例の形式		
1行の長さ	最大54字(うち空白4字、 最大25単語の文脈が入る)	最大43字(うち空白11字、 最大12単語の文脈が入る)
第1キーの位置	左(あるいは右)から 15字目	左(あるいは右)から 8字目
サンプリング	ソート済み切り出し列の集合から通し番号順に $\frac{1}{5}$ 均等抽出	5単語列の異なりごとに先頭の1文例を抽出

4.2 長単位内短単位列の書き換え

4.1に示す相違点のうち、(2)の長単位内短単位列の書き換えについて少し詳しく述べてみよう。

切り出し列を長くとりたい、名詞の連続等は1品詞に書き換えてはどうか。これがこの作業の出発点であった。ただし、長単位間の品詞列も書き換えると、

---- この名探いずれも ---- のようなものもあるので書き換えは長単位内短単位列を
(名) (名)

対象とした。図5辞書テープの形式のBが複数個のものが書き換え対象となる。

○ 長単位内短単位列の異なりパターンと頻度

まず、対象となるもののパターンを検討した。表4がパターン別頻度の一部で

KOSUU

PATTERN

あり、図7がパターン別例である。

表4

* ZISYO TAPE NO TANTANI NO PATTERN *

1	1111116
6	111111
51	11111
2	111161
32	11116
1	11117
298	1111
1	1111Y1
1	11121
8	1112
4	111611
16	11161
1	111666
9	11166
4	111676
236	1116
1	1116P
1	1116R1
1	1116X611
1	1116X6611
2	1116X6
1	1116Y
1	111711
1	11176
1	1117
1887	111

電話(1) 土木(1) 技術(1) 者(6)
 都(1) 計量(1) 検査(1) 所長(6)
 都(1) 教組(1) 委員(1) 長(6)
 都立(1) 交通(1) 研究(1) 所(6)
 国連(1) 中心(1) 主(1) 的(6) な(P)
 娘(1) 芝居(1) 離(1) 記(6) の(R) 巻(1)
 フランデ(1) 協奏(1) 曲(1) 第(6) 4(X) 番(6) ト(1) 長調(1)
 バイオリン(1) 協奏(1) 曲(1) 第(6) 1(X) 番(6) 変(6) ロ(1) 長調(1)
 オルガン(1) 協奏(1) 曲(1) 第(6) 4(X) 番(6)
 オピアノ(1) 協奏(1) 曲(1) 第(6) 2(X) 番(6)
 化学(1) 反応(1) 速度(1) 論(6) (0116) 二〇(8)
 新人(1) 会(1) 創立(1) 五(7) 周年(1) 記念(1)
 ヒアノ(1) 協奏(1) 曲(1) 一(7) 番(6)
 歌謡(1) 曲(1) ベスト(7) テン(7)
 動統(1) 年数(1) 別(1)

図7 パターン別例

参考のために、上位の短単位列を書き出してみると、

1	名詞・名詞	10549
2	名詞・接辞	5484
3	固有名詞・名詞	4189
4	固有名詞・固有名詞	4066
5	サ変語幹・動詞	2915

という具合である。

○ 規則作成と適用

上のパターン表と頻度と例を充分に検討した上で、書き換えすべき範囲を決め規則を作成した。書き換えすべき範囲は基本的に複合語レベルのものとした。規則の一部が表5にある。規則数140、規則の一般形は $a_1 \dots a_m \rightarrow b_1 \dots b_n$ ($2 \leq m \leq 5$, $1 \leq n \leq 2$) である。左の規則を所定の順に適用して書き換え

を行なった結果は表6 図8に示す通りである。

711	1	0010
1167	1	0021
116XX	1	0022
11XX	1	0023
11X	1	0024
117	1	0025
11	1	0020
22	1	0030
66	6	0040
77	7	0050
88	8	0060
CC	C	0061
EE	E	0070
RR	R	0080
XX	X	0090
VY	Y	0100
81176	1	0101
6XX11	8	0102
6XX16	8	0103
7V771	1	0104
1776	7	0105
1X61	1	0106
1X66	1	0107
616E	E	0108
616P	4P	0109
6X11	8	0110
6X16	8	0111
7161	1	0112
7171	1	0120
762E	E	0121
8YXX	X	0122
X1X1	1	0123
X6V1	1	0124
12E	E	0125
16E	E	0126
16P	4P	0127

表5 規則

* TANTANI NO KECUGOO NO YOOSU (OGINO) *		
曲芸(1)	師(6)	曲芸師(1)
曲射(1)	砲(6)	曲射砲(1)
曲(1)	独(6)	曲独(1)
計画(1)	案(6)	計画案(1)
計画(1)	通(6) 長(6)	計画通(1) 計画局長(1)
計画(1)	通(6) だっ(P)	計画通(1) 計画通り(1) だっ(P)
計画(1)	的(6) な(P)	計画的(1) 計画的(1) な(P)
計画(1)	的(6) に(P)	計画的(1) 計画的(1) に(P)
計画(1)	内容(1)	計画内容(1)
計画(1)	部(1)	計画部(1)
計画(1)	作成(1)	計画作成(1)
計画(1)	者(6)	計画者(1)
計画(1)	立案(1)	計画立案(1)
計画(1)	さ(E)	計画さ(E)
計画(1)	し(E)	計画し(E)
計画(1)	どおり(6)	計画どおり(1)
計画(1)	追船(1)	計画追船(1)
計画(1)	行番(1)	計画行番(1)
計算(1)	通(2)	計算通(1)
計算(1)	機(6) 数学(1) 夜話(2)	計算機数学夜話(1)

図8 書き換え例

表6 パターン別書き換えの結果

5 出力結果の数量的検討

4 で述べた作業の相違が出力結果にどのように影響しているか、数字の上でみてみよう。

○ 頻度表による比較

(品詞列) -表 6-

	前 回			今 回		
	品詞	度数	割合	品詞	度数	割合
1	名詞	302603	30.7	助詞	99625	27.2
2	数字	141659	14.4	名詞	98015	26.7
3	格助詞	99302	10.1	動詞	40679	11.1
4	記号	85798	8.7	助動詞	23959	6.5
5	接辞	71483	7.3	読点	22772	6.2
6	動詞	51167	5.2	区切り	16653	4.5
7	数詞	42415	4.3	記号	14574	4.0
8	区切り	35123	3.6	句点	11683	3.2
9	読点	30083	3.1	数詞	11620	3.2
10	助動詞	29704	3.0	副詞	5651	1.5
11	係助詞	22584	2.3	形容詞	4169	1.1

数字・記号・接辞が減少している。これは作業相違利 (1) 広告文等の削除 (2) 短単位列の書き換えによる。助詞が1位になったのは (4) 品詞づけの変更により係助詞・格助詞・終助詞を一つにまとめたことと、(2) により「名詞」の連続がまとめられ「名詞」の頻度が減ったためと思われる。

(2 品詞列) -表 7-

	前 回			今 回		
	品詞列	度数	割合	品詞列	度数	割合
1	名詞・名詞	104472	10.6	名詞・助詞	66085	18.0
2	数字・数字	85640	8.7	助詞・名詞	35969	9.8
3	名詞・格助詞	65079	6.6	助詞・動詞	34569	9.4
4	格助詞・名詞	39706	4.0	読点・名詞	15847	4.3
5	記号・名詞	37722	3.8	動詞・助動詞	13939	3.8
6	名詞・接辞	36529	3.7	句点・区切り	12735	3.5
7	名詞・記号	31718	3.2	助詞・読点	10872	3.0
8	格助詞・動詞	26649	2.7	区切り・名詞	9980	2.7
9	記号・数字	24698	2.5	名詞・名詞	8688	2.4
10	接辞・名詞	24251	2.5	記号・名詞	8263	2.3

前回の「名詞」「名詞」、「名詞」×「接辞」、「数字」×「数字」、「記号」×「名詞」等上位に入っているものが (2) 短単位列の書き換えによりまとめられ、今回の上位列は日本語の構文的な並びを示しているようである。

(3 品詞列) -表 8-

	前 回			今 回		
	品詞列	度数	割合	品詞列	度数	割合
1	数字・数字・数字	45746	4.6	名詞・助詞・名詞	27334	7.5
2	名詞・名詞・名詞	43672	4.4	助詞・名詞・助詞	26452	7.2
3	名詞・格助詞・名詞	27869	2.8	名詞・助詞・動詞	21950	6.0
4	名詞・格助詞・動詞	18670	1.9	助詞・動詞・助動詞	11604	3.2
5	数字・数字・記号	17235	1.7	読点・名詞・助詞	11088	3.0
6	名詞・記号・名詞	17210	1.7	助詞・動詞・助詞	10919	3.0
7	記号・数字・数字	16482	1.7	助詞・読点・名詞	6316	1.7
8	格助詞・名詞・格助詞	16126	1.6	区切り・名詞・助詞	5919	1.6
9	名詞・格助詞・格助詞	15647	1.6	動詞・助詞・動詞	5639	1.5
10	数字・記号・数字	15194	1.5	句点・区切り・名詞	5355	1.5

2 品詞列でみた場合とほぼ同じことが言える。

6 利用例

・ 頻度表による日本語文作成

3品詞列の頻度順表を使って、日本語の品詞の連続の典型を作ってみよう。
文を切りから出発して、その後何がいやすいかを順にみていく。ここでは一番
頻度の高いものをもって、かつ一度とったパターンの場合には次の頻度のものをも
っていく。

(左順3品詞列頻度表より) 井 名 助 名 助 動 陳 句 井

(右順3品詞列頻度表より) 井 名 助 読 名 助 名 助 動 陳 句 井

もっともらしい日本語文の品詞の連続ができていった。このような検討を繰り
返していくつかの文パターンを作ることも可能であろう。

左順3品詞列頻度表 表9

②名助名	27334	③助名助	26452	④名助助	21950	⑤助助陳	11604	⑥助名助	11088	助助助	10919
助助名	6316	①#名助	5919	動助動	5639	句#名	5355	助助名	5297	⑦陳句#	5209
名助名	4822	動名助	4513	陳名助	4423	名名助	4332	名助助	4228	⑧動陳句	3571

右順3品詞列頻度表 表10

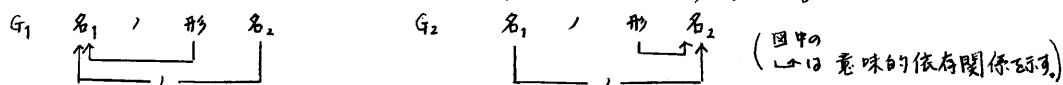
④名助名	27334	⑤助名助	26452	④名助助	21950	③助助陳	11604	⑦助名助	11088	助助助	10919
⑧助助名	6316	①#名助	5919	動助動	5639	句#名	5355	助助名	5297	①陳句#	5209
名助名	4822	動名助	4513	陳名助	4423	名名助	4332	④名助助	4228	②動陳句	3571
動陳名	3362	動陳助	3355	動句#	3258	助動句	3146	記名記	2918	名記名	2813

・ 「名詞関係依存の原則」考察

田中らが先に発表した⁹⁾上記原則をこの「品詞列集成」を使って考察してみた。
これは、 $\langle 名_1 \times 助 \rangle$ 、 $\langle 形 \times 名_2 \rangle$ の二つの $\langle 名 \rangle$ の意味的依存関係から $\langle 形 \rangle$ の
修飾先を決定するというものである。

(1) $\langle 名_2 \rangle / \langle 名_1 \rangle$ が成立すれば、 $\langle 形 \rangle$ は前の $\langle 名_1 \rangle$ を修飾する。— G_1 の型—

(2) $\langle 名_1 \rangle / \langle 名_2 \rangle$ が成立すれば、 $\langle 形 \rangle$ は後の $\langle 名_2 \rangle$ を修飾する。— G_2 の型—



上記原則について、「品詞列集成」を使って調べたところ反例なしで一般性が確
認された。図1に示すように、 $\langle 名 \times 助 \rangle$ 、 $\langle 形 \rangle$ の部分から $\langle 名 \rangle$ 、 $\langle 助 \rangle$ 、 $\langle 形 \rangle$ 、 $\langle 名 \rangle$ の列を
探し1例ずつ検討した。考察結果の詳細と文例は文献⁹⁾にもある。

7 問題点

・ 品詞づけ

原ラフあるいは原ラフ加工の段階で品詞づけの誤りが生じている。

例え	誤り品詞	正しい品詞
まい	名詞 →	助動詞
この	名詞 →	連体詞
なら	動詞 →	動詞か助動詞

等である。これらは、前回の品詞列集成から見つけたもので原ラフの誤りと思
われる。今回作業以前に見つけられたもので頻度の多いものについては、部分的
な修正を行なったが、その後見つけられたもので

では 接続詞 → 助詞 か 助動詞 + 助詞

等は未修正である。

また、長単位内短単位列書を換えて、助詞を含まない文相当の短単位列、

潜水艦 潜航せず
1 3 E P

のようなものに短単位列書を換之規則が適用され、誤っておとめられ品詞づけされたものもある。

以上のような内容で、誤り品詞からなる単語列も含まれているので、利用者は注意してほしい。

最後に、原データを御提供下さった国立国語研究所、並びに前回作業を手がけられた西村恕彦氏（東京農工大）、今回作業を御援助下さった洲一博パターン情報部長に謝意を表する。

文献

- 1) 国立国語研究所：“電子計算機による新聞の語彙調査”，国研報告37，1970
- 2) 3) 田中穂積，萩野為野，萩野綱男：“新編日本語全品詞列集成 左順 右順”，電総研推論機構研究室，1979
- 4) 5) 同上：“新編日本語品詞列集成 左順 右順”，同上，1979
- 6) 7) 西村恕彦，萩野綱男：“日本語品詞列集成 左順 右順”，電総研，1977
- 8) 田中穂積，萩野為野：“形容詞もしくは形容動詞の連体句と含む名詞句の意味処理”，情報処理学会大会予稿，昭和54年
- 9) 同上：“形容詞もしくは形容動詞の修飾先の名詞を決める原則について”，文部省特定研究報告，昭和54年