

リント上に現われなく、後述する配列等の問題を解決するため、表1に示す符号変換を行う。

第1表 漢テレ-HITACコード変換表

漢テレコード		HITAC コード	
0	128	ホ	163
1	001	マ	164
2	002	ミ	165
3	131	ム	166
4	004	メ	167
5	133	エ	168
6	134	ヤ	169
7	007	ユ	170
8	008	ヨ	172
9	137	ラ	173
ち	138	リ	174
=	011	ル	175
↓	140	レ	186
+	016	ロ	187
A	145	ワ	188
B	146	ン	189
C	019	A	193
D	148	B	194
E	021	C	195
F	022	D	196
G	151	E	197
H	152	F	198
I	025	G	199
?	026	H	200
・	155	I	201
)	028	J	209
-	032	K	210
J	161	L	211
K	162	M	212
L	035	N	213

漢テレコード		HITAC コード	
M	164	O	214
N	037	P	215
O	038	Q	216
P	167	R	217
Q	168	S	226
R	041	T	227
!	042	U	228
¥	171	V	229
*	044	W	230
∅	176	X	231
/	049	Y	232
S	050	Z	233
T	179	0	240
U	052	1	241
V	181	2	242
W	182	3	243
X	055	4	244
Y	056	5	245
Z	185	6	246
←	186	7	247
ゝ	059	8	248
(	188	9	249
(161/164)	013		064
	064		064
上記以外のコード		ゝ	095

データの構造 入力データの構造については、KWIC Index File の場合の入力形式と同じであるが、1回の処理の対象となるデータをファイルと定義する。漢テレによるデータの打鍵のとき、表2に示す記号を pre editing として挿入

する。

実際の原文中には、つぎのような問題点がある。

- 1) 強調文(例: ニンで述べらる重宝とは、哲学的なものである)。
- 2) 文の途中の図、化学式、表、数式およびその説明文。
- 3) 箇条書き文。
- 4) ページ
- 5) 参考文献の指標

本来、原文が有するあらゆる情報を翻字すべきであるが、実際にはその処理目的への必要性和翻字化の煩雑さとの兼ね合いで決定されるものであろう。今回は、本報告に使用したデータを中心に以下の処理を施した。

図およびその説明文は、すべて除外し、数式は可能な限り原データに忠実に入れる。なお、数式の終わりには、これを文と見立て、文末の記号「。」を挿入した。(「。」がなければ、数式と次の文がつながったものと計算機は解釈する。) ページは注釈とみなし処理した。

参考文献の指標は、原データで「...」である。「2)」と「...」の部分には「...」である「2)」と変更した。文の処理の判定で、「2)」をつづいた文と解釈されるためである。

なお、漢テレにない未登録の漢字については、表2表に示すように、その読みをカタカナに直して(カタ)でその表現方法を用いた。

例: 両機類 → 両リセイリ類

翻字の手順は、

- 1) 原データを入力構造の拘束に従って Pre Editing を行った。
- 2) Pre Editing されたファイルを漢テレを使って紙テープに穿孔する。
- 3) 紙テープを計算機により磁気テープ(MT)に転写する。

3. 分ち書きの手法

日本語には、英語文にみられるような語を単位に分ち書きせよと記述されるという特徴がある。本システムの基本ルーチンである ALPS-KNDC Index File は、分ち書きされた語を単位とする処理方法が用いられているため、Pre Editing として、人手によるか計算機による分ち書き処理を必要とする。

第2表 入力データ構造

<ファイル>	::=<データ>▷
<データ>	::=<テキスト> <データ><テキスト>
<テキスト>	::=<テキスト名><注釈><著書名><テキスト内容>
<テキスト名>	::=<文>
<注釈>	::=<文>
<著書名>	::=<文>
<テキスト内容>	::=<章> <テキスト内容><章>
<章>	::=<章名><章内容>
<章名>	::=<文>
<章内容>	::=<節> <章内容><節>
<節>	::=<節名> <節内容>
<節名>	::=<文>
<節内容>	::=<段> <節内容><段>
<段>	::=<文> <段><文>
<文>	::=<文内容>
<文内容>	::=<語> <文内容><語>
<語>	::=<文字> <語><文字>
<文字>	::=<外国文字> <カタカナ> <ひらがな> <漢字> <数字> <特殊文字> <漢テレにない字>
<外国文字>	::=ABC...abc...AaBb...αβγ...
<カタカナ>	::=アイウ...
<ひらがな>	::=あいう... ちゃゅよ...
<漢字>	::= ㇿ々亜阿... 梓濟腕
<数字>	::= 1 2 3 ... 1 2 3 ... 1) 2) 3)
<特殊文字>	::= + - / Δ * ...
<漢テレにない字>	::= <カタカナ>

但し、上に用いられた記号は次のことを意味する。

<	>	; 最小単位の集合
::		; 左辺は右辺の構造で成り立つ
		; 単位の和集合
<> <> <>		; 繰り返しを許した和集合
...		; 同種類の文字によって左辺の構造を規定する

自動分ち書きについては、多くの方法が紹介されている。ここでは、植村氏が報告されている単純な手続きによる方法を採用した。すなわち、文字種の異りによる分割という方法である。漢テレのコード配分は、字種に対して部分集合を形成するように設計されている。とくに1文字コード(1および2列)中のオ1列のコードも判定するだけで、字種識別が可能で、単純で高速な処理ができる。その処理手順は、入力される連続した文字列中、隣接した2文字を比較し、その字種の異なった部分にスペースを挿入する。ただし、未登録の文字であるデータポイントには含まれたカナの部分については、これを漢字種と判定し、分割を行わない。

処理結果 日本語においては正書法の規則がヨーロッパの諸言語に比較して厳密に守られていないという問題がある。例2図の例で示すように、用言の語尾、名詞、用言の漢字かな混り表現、数の表現が分離する例と副詞、接続詞等にみられる分離が行なわれない例が多くみられた。後者については、量的に研究を用いて設定が可能であるが、前者の場合には、多くの問題を含んでいる。

- (a) 見立て → 見立_レて 目覚まし → 目覚_レまし
 かけ離れた → かけ_レ離れた 抽き出す → 抽_レき出_ス
 は虫類 → は_レ虫類
 オ1図 → オ_レ1_レ図 20億年 → 20_レ億年
- (b) 一見_レ臆_レが → 一見_レ臆_レが
 一方_レ抑制 → 一方_レ抑制
 論じた_レもっとも → 論_レじたもっとも
例2図 分ち書きの不成功例

4. 記事の作成

作成されたコンコーダンスの利用にあたっては、その Keyword を含む文脈の原データとの参照が必要となる。その記事の書式を2種類とした。

- 1) 例3図(a)で示すように、原データでのページ、行を人手により、漢テレで翻字するとともに挿入する。
- 2) (b)に示すように、計算機で自動的に、文番号と同文中での行番号を挿入する。

この2つの方法が選択できる。

(a)

原データ

/ 通 000 / _レ

/ _レ 01 _レ 本 文

/ _レ 02 _レ _____

/ _レ 03 _レ _____

/ _レ 03 _レ _____

(b)

/ _レ 本 文

/ _レ _____

/ _レ _____

原データの書式

_レ 01 _レ _レ 02 _レ _レ 03 _レ 0004 _レ 05 _レ

(表題番号) (章番号) (節番号) (頁番号) (行番号)

_レ 01 _レ _レ 02 _レ _レ 03 _レ 0004 _レ 05 _レ

(表題番号) (章番号) (節番号) (文番号) (行番号)

参照番号欄の書式

_レ : スペース

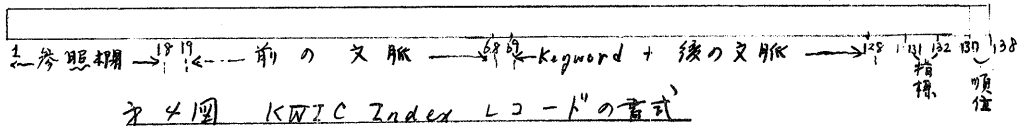
例3図 入力データの書式

文の1行の長さは、出力装置によって調整が可能であり、語間のスペースは、1行に標準化が行われている。

作成の手順は、入力データより、指標判別を行ない、参照番号をセットする。1行分の文脈を移した後、MTに蓄積される。なお、KWIC Indexの対称から除外した表題、注等については、参照番号欄の行数、文数を0とする。

5. KWIC Index Fileの作成手順

KWIC Indexの作成は、記事のMTを入力として、文を単位にコピーに蓄積される。この文からKeywordの認定を行ない、参照番号のセット、KeywordとそのKeywordに続く前後の文脈を移し、おたの書式でMTに出力する。これを1レコードとする。



すなわち、

(1) 参照番号: 記事中のKeywordを含む文脈に与えられている参照番号がそのまゝ与えられる。

(2) 前の文脈: 記事中のKeywordの前の文脈を規定数だけ出力させる。ただしその文の先頭が規定字数内にある場合は、それ以前をスペースとする。

(3) 後の文脈: 記事中のKeywordを含む後の文脈を規定字数だけ出力する。ただし、規定字数内で、文が終っている場合は、残余をスペースとする。

(4) 指標: 記事中、標題、章名、節名の指標、Keywordが含まれている行にコメントが付されている場合、右端にこれを示すこととする。

(5) 出現順位: 記事中に、そのKeywordが出現した順位を示す。

なお、逆配列(語尾から語頭へ文字配列し、アルファベットあるいは五十音順にソートする)のKWIC Index Fileの作成は、1レコードを中心で反転したレコードについて、ソートを行ない、その結果を逆転して出力する。

1ファイル分のKWIC IndexレコードがMTに蓄積されると、COBOLのソート・プログラムを用い、ディスク・ソートを行なう。すなわち、Keywordとそれに続く文脈について、指定の字数についてソートされる。結果は、記事の後に続き、MTにKWIC Index Fileとして蓄積される。

6. Keywordの統計処理

KWIC Index Fileで得られた結果を用い、自然言語の分析に必要な以下に示すいくつかの統計データを計算し出力するプログラムを作成した。

(1) 延べ語数: 記事中に出現したKeywordの総数。

(2) 異なり語数: 記事中に出現した異なったKeywordの総数。

(3) 異なり語の出現頻度数。

(4) 延べ語数に対する各異なり語の出現頻度数の割合。

(5) (4)の累積頻度。

処理手順は、出力形式により、Keyword配列と頻度順配列の2つがある。

Keyword配列

(1) Keywordの総数を記憶。

- (2) KWIC Index File からレコード単位で読み込み Keyword の抽出を行なう。
- (3) 前の Keyword と比較、同じ場合は出現頻度の計算、異なっている場合は、前の Keyword についての情報を MT に出力し、新しい Keyword をセットする。
- (4) (2), (3) を繰返し、全レコードの処理を終ると、お5図に示す統計情報を出力する。

```

***** 統計情報 *****
*
* 文献名 脳とオートマトン。
*
* 見出し語の取り出し方法 逆順
*
* 見出し語の配列方法 出現頻度順
*
* 異なり語数 1116語
*
* 総語数 3425語
*
*****

```

頻度順配列

お5図 統計情報の出力例

- (1) Keyword 配列の MT を入力とし、出現頻度数をお1 Key, Keyword をお2 Key として、ソートを行なう。
- (2) (1) の結果について、累積頻度を計算し出力する。
- (3) (1), (2) を繰返し、全レコードを処理後、統計情報を出力する。
- 以上の手続を行なうて、Keyword の取だし方法(正, 逆), 配列順序(Keyword, 頻度)の4種のうち、必要なものを選択出力できる。

7. 出力形式(たて横変換)

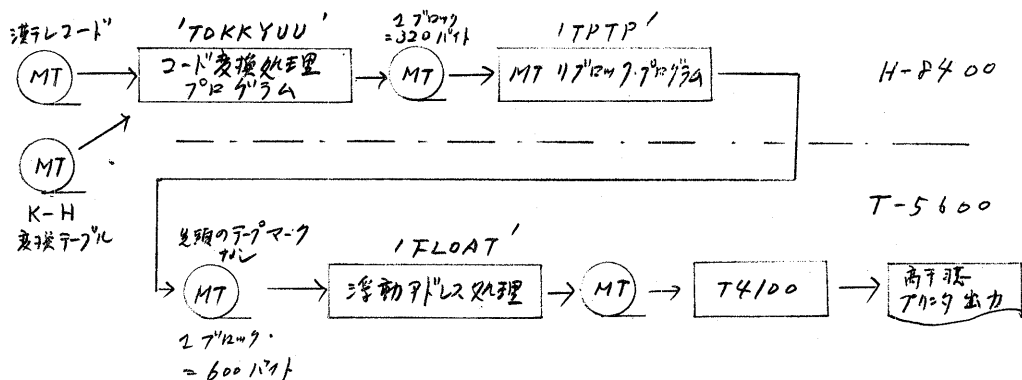
KWIC Index を印刷するには、漢字レの場合も、後述する高速度プリンタでも、紙の横幅に物理的な制限があり、MT 上に蓄積されている1レコードを全部印字することができない。そこで以下に示すいくつかの方法がある。

- (1) KWIC Index レコードを左右に2分し、別々に印刷し、後に貼付して横書きにする。
- (2) 前文脈の左端から後文脈の右端を切りとり中心部分30字程度を印字し、横書きにする。
- (3) 横書きをたて書きに変換して、レコードの全文字を出力する。
- (1) の方法は、漢字レの行送り機構の精度が低いため貼付の左右でずれが生じる。また(2)の方法は、文脈が短くなり、KWIC の特徴を損うといった欠点があり、たての一行の長さも自由に選べる(漢字レ)(3)の方法もこゝでは採用した。この方法の欠点は、特殊記号、数字等たて書き用の記号がない場合、欧文のみ小さく等完全な変換が行なえない点がある。

処理手順は、1頁(40行)分づつコア内に読み込み、参照番号を漢数字に変換して下段にたて書きとし、前の文脈を上段、Keyword と後の文脈を下段に漢字レコードに変換して移す。また頁ごとに、製本に便利のように奇数は左、偶数は右側の下方に頁数がつけられる。これを1単位として MT に出力される。

漢字レで印字する場合には、HITAC の400のユーザイリタイプログラムにより紙テープに出力し、漢字レで打鍵させる。(打鍵速度 = 130字/分)

高速度プリンタ・システム(AT4/00)を用いた高速印刷を行なう場合、お6図に示すように、コード変換処理と MT プロッタ・フォーマットの調整を必要とする。



第6図 T4100 処理手順

30

コード変換処理プログラム（プログラム名 'TOKKYUU'）は植村氏作成された漢字コード間の相互変換プログラムで、ここでは、漢字からT4100コードへの変換処理を行なった。

浮動アドレス処理サポートプログラム 'FLOAT' は、T4100 システムの浮動アドレス方式プリントの入力 MT を作製する目的で当所の推論機構研究室で開発され、TOSBAC-5600 を使用するユーザに開放されてこのプログラムである。

得られた MT を T4100 システムへかけることにより、記号、KWIC Index File、統計処理結果等が高速印字出力される。（1文字 = 32 × 32 ビット、字種 = 8192）

8. 処理結果と応用

現在、数万語からなる標本 6 種について出力結果が得られており、さらに 20 万字からなる特許公報のコーンコードエスを作成中である。また、T4100 に 2 文字 10 ビットで記憶させることになり、数万語からなる百代に 2 文字のコーンコードエスを作成中である。

また、KWIC Index File に記憶させてある Keyword の出現順位と、異なり語配列表を利用して、高速な辞書検索システム、逆配列表による活用処理実験、頻度配列表による電算 Keyword の抽出等の応用を考えている。

最後に、本システムの仮々のプログラム作成を手伝って下さった、電気通信大学の谷津直和、佐藤雅之、石田富士夫の諸君に感謝いたします。

参考文献 (1) 坂本義行: Automatic Language Processing System, KWIC Index File, CL 研究委員会, 69-5, 1969.3.

(2) 坂本義行, 岡本哲也, 北藤知巳: ALPS-KWIC Index File の応用 - Concordance と Syntax KWIC の作成, 第 7 回情報科学研究集会発表論文集, JICST, 1970.

(3) 西村勉彦: 電算漢字仕様案, 機械翻訳研究委員会, 1965.9.

(4) 植村俊亮: 電子計算機による自動索引の研究 (上), 電子技術総合研究所研究報告 第 743 号, P60, 1974.2.

付録 I 記事と KWIC Index の出力例

数理科学 NO. 98, AUGUST 1971 Page 24~30

[illegible]

PANN

[illegible][illegible]

付 録 Ⅱ 正配列の統計出力例

脳とオートマトン。

頻度表 (配列順) 正順		18頁		
順位	見出し語	頻度数	相対頻度	累積頻度
8551	上段	1	0.00003	0.82242
8552	成功	1	0.00003	0.82245
8553	条件	2	0.00006	0.82251
8554	状	1	0.00003	0.82254
8555	状態	4	0.00012	0.82266
8556	定式化	1	0.00003	0.82269
8557	情報交換	1	0.00003	0.82272
8558	情報処理	55	0.00155	0.82277
8559	情報処理能力	2	0.00006	0.82283
8560	場	0	0.00000	0.82283
8561	場合	3	0.00009	0.82292
8562	場所	2	0.00006	0.82298
8563	鍵	8	0.00023	0.82321
8564	信号	1	0.00003	0.82324
8565	神経回路	1	0.00003	0.82327
8567	神経回路群	7	0.00020	0.82347
8568	神経回路網	1	0.00003	0.82350
8569	神経回路網内部	3	0.00009	0.82359
8570	神経系	1	0.00003	0.82362
8571	神経細胞	2	0.00006	0.82368
8572	神経生理学者	4	0.00012	0.82380
8573	神経信号	10	0.00029	0.82409
8574	神経網	1	0.00003	0.82412
8575	振動	1	0.00003	0.82415
8576	振幅	1	0.00003	0.82418
8577	進	1	0.00003	0.82421
8578	進化	10	0.00029	0.82450
8579	進歩	1	0.00003	0.82453
8580	新	1	0.00003	0.82456
8581	人間	2	0.00006	0.82462
8582	人工知能	1	0.00003	0.82465
8583	水中生活	1	0.00003	0.82468
8584	推測	2	0.00006	0.82474
8585	誰	1	0.00003	0.82477
8586	…体細胞	2	0.00006	0.82483
8587	数	8	0.00023	0.82506
8588	数学	1	0.00003	0.82509
8589	数学者	1	0.00003	0.82512
8590	数秒	1	0.00003	0.82515
8591	数理言語学	1	0.00003	0.82518
8592	星状細胞	3	0.00009	0.82527
8593	盛	1	0.00003	0.82530
8594	精度	2	0.00006	0.82536
8595	静止電位	2	0.00006	0.82542
8596	整数	1	0.00003	0.82545
8597	示候	1	0.00003	0.82548
8598	接触部	2	0.00006	0.82554
8599	接続	3	0.00009	0.82563
9000	設計	1	0.00003	0.82566

付 録 Ⅲ 逆配列の頻度順統計出力例

脳とオートマトン。

順位	頻度表 (頻度順) 逆順		頻度数	相対頻度	1 頁 累積頻度
	-----	見出し語 -----			
1		の	274	0.0800	0.0800
2		ニューロンに	237	0.0692	0.1492
3		をば()で	104	0.0304	0.1796
4		シナプス	83	0.0242	0.2038
5		脳と	77	0.0225	0.2263
6		その	75	0.0219	0.2482
7		も	74	0.0216	0.2698
8		神経回路網	64	0.0187	0.2885
9		な	48	0.0140	0.3025
10		図	41	0.0128	0.3153
11		小脳	40	0.0127	0.3280
12		の	34	0.0097	0.3377
13		こ	24	0.0067	0.3444
14		す	22	0.0064	0.3508
15		る	21	0.0061	0.3569
16		電	20	0.0058	0.3627
17		位	20	0.0058	0.3685
18		0	17	0.0050	0.3735
19		して	16	0.0047	0.3782
20		は	16	0.0047	0.3829
21		管	16	0.0047	0.3876
22		と	16	0.0047	0.3923
23		に	16	0.0047	0.3970
24		興	16	0.0047	0.4017
25		有	15	0.0044	0.4061
26		大	15	0.0044	0.4105
27		脳	14	0.0041	0.4146
28		細胞	14	0.0041	0.4187
29		という	13	0.0038	0.4225
30		な	12	0.0035	0.4260
31		化	12	0.0035	0.4295
32		多	12	0.0035	0.4330
33		量	12	0.0035	0.4365
34		で	11	0.0032	0.4397
35		大	11	0.0032	0.4429
36		脳	11	0.0032	0.4461
37		細胞	11	0.0032	0.4493
38		と	11	0.0032	0.4525
39		いう	10	0.0029	0.4554
40		な	10	0.0029	0.4583
41		化	10	0.0029	0.4612
42		多	10	0.0029	0.4641
43		量	10	0.0029	0.4670
44		で	10	0.0029	0.4699
45		大	9	0.0026	0.4725
46		脳	9	0.0026	0.4751
47		細胞	9	0.0026	0.4777
48		という	9	0.0026	0.4803
49		な	9	0.0026	0.4829
50		化	9	0.0026	0.4855