

# コンピューターによる

## 私用コンコーダンスの編纂

松 尾 雅 嗣

### は じ め に

文学，語学研究におけるコンコーダンスの利用価値を否定する人はないであろうし，また今日コンコーダンスの編纂におけるコンピュータの有用性を否定する人もないであろう。しかし，文学，語学，あるいは広い意味での言語データ，の研究者にとって，自分の研究対象である，作家，作品，あるいは言語データのコンコーダンスが公刊されていない場合，自力でコンピュータを利用してコンコーダンスを作することは，実際にはともかく，心理的には，不可能に近いと言ってよいであろう。

本稿は，現在開発中のプログラム・パッケージLEXを用いることにより，コンコーダンスの作成が，容易にしかも短時間に，少なくとも手作業に比べれば，格段に容易に短時間に，行なえることを具体例によって示すことを目的とする。

LEXは，コンピュータによるテキストの語彙処理を目的として開発中のプログラム・パッケージであり，人文科学，社会科学の研究者も容易に利用できることを特に意図したものである。システムとしてのLEXには，頻度順リスト，アルファベット順リスト（アイウエオ順リスト），索引，脚韻語索引などを作成する機能があり，そのひとつとして，コンコーダンスを作成する機能も備わっている。この機能を利用することにより，研究者は容易にコンコーダンスを編纂することができる。そして，その際，LEXに備わっているデータの選別機能を使うことによって，作成されるコンコーダンスを研究者個人の研究目的に応じて編集することも，即ち「私用」のコンコーダンスを編纂することも，一定の範囲内ではあるが，可能である。

なお，LEXの全体像および詳細に関しては，本稿では割愛せざるをえないので，文献(1)，(2)，(3)を参照されたい。

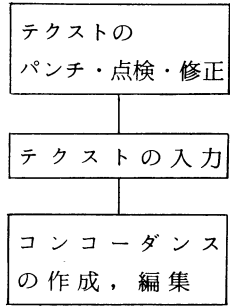
### 1. データの作成とテキストの入力

LEXによってコンコーダンスを作成する作業の手順は次の図1に示す3ステップに分けられる。

コンコーダンスを作成するためには，図1に示すように，まずテキストをカードにパンチしなければならない。パンチしたカードの点検，パンチミスの修正も含めて，この作業が最も時間と労力を必要とする。

テキストをカードにパンチする時注意すべき点をここで簡単に述べておこう。まず第一に、コンコーダンス系の出力で1行として印刷したいデータ、例えば、原文の1行、1文などを、必ずカード1枚にパンチする。このカード1枚分のデータが1行として印刷されることになる。次に、データ中の単語は、どのような文字、記号から成っていてもよいが、長さは16文字以内でなくてはならない。但し、記号については、次に挙げる各記号は単語の区切り等として用いられるので単語の構成要素とはなりえない。

図1



空 白 (スペース) 疑 問 符 ?  
 コ ロ ン : 感 嘆 符 !  
 セミコロン ; 引 用 符 ▽  
 ピリオド ・ カ ン マ ,  
 左 か っ こ ( 右 か っ こ )  
 縦 線 (ストローク) | 下 線 \_

以上12個の記号は、単語の区切りとして用いられるが、コンコーダンス系の出力では入力したままの形で印刷される。また次の4個の記号は、行、頁(単語連鎖と称する)の区切りとして用いられるので、特別の必要がない限り用いてはならない。

星 印 (アステリスク) ※ パーセント記号 %  
 シ ャ ー プ # 単 価 記 号 @

次の図2に示すのが、実際にカードにパンチしたデータをそのままの形で印刷したものである。

図2で、各行の左の数字は参照の便のために付したもので、実際のカードにはパンチされていない。各々の数字の右側がそれぞれカード1枚にパンチされた内容である。

①～⑤は、テキストをシステムに入力するための指示である。LEXでは、利用者<sup>1)</sup>とシステムとのコミュニケーションは、LEXコマンドと、LEX文によって行なわれる。この場合、①～⑤がLEX文であり、実際にテキストを入力する時には、LEX文の前に、

// LEX DSP=NEW

というLEXコマンドが必要である。

LEX文のうち、①はINPUT 命令で、テキストをシステムに入力せよという命令である。②は註釈ステートメントで、“<COMMENT>”で始まるカードは、LEX文中の註釈のためにのみ用いられ、データの処理にはまったく関係しない。③～⑤は、INPUT 命令に対する明細指示である。④では、⑦から始まるテキストの各行を“LINE”と呼び、3921から始まる一連番号を各行につけることを意味する。もし、“START=3921”がなければ、テキストの各行には、省略時解釈により、1から始まる一連番号が付けられる。

カード番号⑥は註釈行である。斜線（ / ）を含むカードは、テキスト中のどこにあっても、テキスト本体とは見做されず、註釈行と見做される。註釈行は言うまでもなくコンコーダンスの出力には無関係である。

カード番号⑦以降がシステムに入力されるテキストの本体である。図2のテキストは、③からも明かなように、Geoffrey Chaucerの*Canterbury Tales* 中の“Reeve's Tale”である。

以上データ入力の実例を挙げ、ごく大ざっぱな説明を加えたが、詳しくは末尾に携げた文献を参照されたい。

## 2 出力の形式

図2に例を示したようなデータをINPUT 命令によって入力すると、入力されたテキストは通常は、計算センターの公用ボリューム上に恒久的ファイルとして記憶され、以後のデータ処理ではその都度データを入力する必要はなくなる。

データが正常に入力されれば、コンコーダンス作成の準備が整ったことになる。LEXでは、コンコーダンス系の出力として、大別して次の2種類の出力形式が可能である。

ひとつは次の図3に示すような通常のコンコーダンス形式の出力である。この場合、単語はアルファベット順で、各語の頻度と、各語が出現する行と、その行番号が出力される。

図3で、各行の右端に、図2のINPUT 命令に続けて、

LS1 NAME=LINE, START= 3921,

という指示を与えた結果として、“LINE”という名称と、各行の行番号が印刷されていることに注意されたい。行番号のはかに識別値や名称をつけたい場合については、文献(2)、(3)を参照されたい。

この形式でテキスト中のすべての単語についてのコンコーダンスを出力したいのであれば、LEX コマンドとLEX文は次のようになる。<sup>2)</sup>

// LEX

CŌNCŌR (またはCŌNCŌRDANCE)

図2 入力データの例

```
1 INPUT
2 <COMMENT> EXAMPLE OF COMMENT
3 TITLE=REEVE'S TALE
4 LS1 NAME=LINE,START=3921,
5 READ DATA
6 / EXAMPLE DATA FOR CONCORDANCE AND KWIC INDEX
7 AT TRUMPYNGTOWN, NAT FER FRO CANTEBRIGGE,
8 THER GOOTH A BROOK, AND OVER THAT A BRIGGE,
9 UPON THE WHICHE BROOK THER STANT A MELLE;
10 AND THIS IS VERRAY SOOTH THAT I YOW TELLE:
11 A MILLERE WAS THER DWELLYNGE MANY A DAY,
12 AS ANY PECOK HE WAS PROUD AND GAY,
13 PIPEN HE KOUDE AND FISSHE, AND NETTES BEETE,
14 AND TURNE COPPES, AND WEL WRASTLE AND SHEETE;
15 BY HIS BELT HE BAAR A LONG PANADE,
16 AND OF A SWERD FUL TRENCHANT WAS THE BLADE,
17 A JOY POPPERE HE IS IN HIS POUCH;
18 THER WAS NO MAN, FOR PERIL, DORST HEYM TOUCHE,
19 A SHEFFELD THWITEL BAAR HE IN HIS HOSE,
20 ROUND WAS HIS FACE, AND CAMUS WAS HIS NOSE;
21 AS PILED AS AN APE WAS HIS SKULLE,
22 HE WAS A MARKET-BETTERE ATTE FULLE,
23 THER DORST NO WIGHT HAND UPON HEYM LEGGE,
24 THAT HE NE SWOR HE SHOLDE ANON ABEGGE,
25 A THEEF HE WAS FOR SOTHE OF CORN AND MELE,
26 AND THAT A SLY, AND USAUNT FOR TO STELE,
27 HIS NAME WAS HOOTE DEYNOUN SYMKYN,
28 A WYF HE HADDE, YCOMEN OF NOBLE KYN;
```

図3 CONCORDANCE 命令の出力例 (部分)

BRYDEL ( 1)	LINE
HE STREPETH OF THE BRYDEL RIGHT ANON.	4003
BRYNGES ( 1)	LINE
SLYK AS HE FYNDES, OR TAA SLYK AS HE BRYNGES.	4130
BURDON ( 1)	LINE
HIS WYF BAR HYM A BURDON, A FUL STRONG;	4165
BUSSHEL ( 3)	LINE
HE HALF A BUSSHEL OF HIR FLOUR MATH TAKE,	4093
THOU SHALT A CAKE OF HALF A BUSSHEL FYNDE	4244
OF HALF A BUSSHEL FLOUR, FUL WEL YBAKE.	4312
BUT ( 25)	LINE
BUT SHE WERE WEL YNORISSED AND A MAYDE,	3948
THER DORSTE NO WIGHT CLEPEN HIRE BUT "DAME";	3956
BUT IF HE WOLDE BE SLAYN OF SYMKYN	3959
BUT RIGHT FAIR WAS HIRE HEER, I WOL NAT LYE.	3976
FOR THERBIFORM HE STAL BUT CURTEISLY,	3997
BUT NOW HE WAS A THEEF OUTRAGEOUSLY,	3998
BUT TEROF SETTE THE MILLERE NAT A TARE;	4000
TO YEVE HEM LEVE, BUT A LITEL STOUNDE,	4007
AND THOUGHT, "AL THIS NYS DOON BUT FOR A WYLE.	4047

図4 KWIC INDEX 命令の出力例 (部分)

BED ( 7)	4139
AND IN HIS OWENE CHAMBRE HEM MADE A BED,	4141
NOGHT FROM HIS OWENE BED TEN FOOT OR TWELVE.	4142
HIS DOUGHTER HADDE A BED, AL BY HIRSELF,	4206
AND I LYE AS DRAF-SAK IN MY BED;	4219
I HADDE ALMOOST GOON TO THE CLERKES BED.	4223
AND FOOND THE BED, AND THOUGHT NOGHT BUT GOOD,	4258
UNTO THE BED THER AS THE MILLERE LAY.	
BEDDE ( 3)	4153
TO BEDDE HE GOTH, AND WITH HYM GOTH HIS WYF.	4159
TO BEDDE WENTE THE DOUGHTER RIGHT ANON;	4160
TO BEDDE GOTH ALEYN AND ALSO JOHN;	
BEDDES ( 2)	4156
THE CRADEL AT HIE BEDDES FEET IS SET,	4213
AND BAAR IT SOFTE UNTO HIS BEDDES FEET.	
BEE ( 1)	4187
BY GODDES SALE, IT SAL NEEN OTHER BEE!.	

これに対して、2 番目の出力形式は、通常KWIC索引と呼ばれる形式である。<sup>3)</sup> 出力例を前の図 4 に示す。

図 4 から明らかなように、KWIC索引の特徴は、エントリーとなる単語が行中のどこにあっても、印刷行中では、常に同じ位置に印字されることである。その他の出力内容については、コンコーダンスの形式とはほぼ同じである。但し、この形式では 1 行の長さを40文字を越える時には、行の一部が印刷されないこともある。1 行が40文字以上の場合のために、次の図 5 に示す形式も用意されているが、出力が多少続みにくくなるきらいがあるし、またこの場合でも、1 行が60文字以上であれば、行の一部が印刷されないこともある。(図 5 では、行番号だけでなく、段落番号も出力されている。)

テキスト中のすべての単語についてのKWIC索引の出力は次のLEXコマンドとLEX文による。<sup>4)</sup>

```
// LEX
KWIC (またはKWIC INDEX)
[HALFSIZE]
```

HALFSIZE を付ければ、図 4 の出力、省略すれば、図 5 の出力が得られる。

### 3. 編集の実例

公刊を目的としない私用のものであれば、2 で述べた方法で出力したもので十分使用に耐えうるであろう。しかし、その際最も問題になるのは同綴異語 (homographs) であろう。ふたつ以上の同綴異語がある時には、それぞれを別個の項目として出力することになる。例えば英語の動詞 “lead” と名詞 “lead” がテキスト中にあり、動詞が、2, 4, 5 行目に、名詞が10行目に現われているとすると、名詞だけを出力するには、

```
// LEX
KWIC
SELECT LSQ
2
4
5
×END
[HALFSIZE]
```

<sup>5)</sup> とすればよい。以下、KWIC INDEX 命令の例だけを示すが、CONCORDANCE 命令の場合もまった

图5 KWIC INDEX 命令出力例(2)部分

ALL ( 4 )  
POLITICAL AND ECONOMIC INTERESTS OF ALL THE NATIONS AND PEOPLES OF THE WORLD AS  
13 1  
DYNAMIC DEVELOPMENT OF DETENTE, ENCOMPASSING ALL SPHERES OF INTERNATIONAL  
21 3  
RELATIONS IN ALL REGIONS OF THE WORLD, WITH THE PARTICIPATION OF ALL COUNT  
22 3  
ONS IN ALL REGIONS OF THE WORLD, WITH THE PARTICIPATION OF ALL COUNTRIES, 22 3  
ALWAYS ( 1 )  
PEACE, HAS ALWAYS BEEN ONE OF THE MOST PROFOUND ASPIRATIONS OF HUMANITY.  
2 1  
AN ( 3 )  
ATTAINMENT OF THE OBJECTIVE OF SECURITY, WHICH IS AN INSEPARABLE ELEMENT OF  
1 1  
AN END TO THIS SITUATION, TO ABANDON THE USE OF FORCE IN INTE  
8 1  
ORTS AIMED AT REACHING THE GOALS OF DEVELOPMENT, TO BECOME AN OBSTACLE ON THE  
18 2  
AND ( 10 )  
AND TO SEEK SECURITY IN DISARMAMENT, THAT IS TO SAY, THROUGH  
9 1  
THE ENDING OF THE ARMS-RACE AND THE ACHIEVEMENT OF REAL DISARMAMENT ARE TASKS OF  
11 1  
PRIMARY IMPORTANCE AND URGENCY. TO MEET THIS HISTORIC CHALLENGE IS IN THE  
12 1  
POLITICAL AND ECONOMIC INTERESTS OF ALL THE NATIONS AND PEOPLES OF THE  
13 1  
POLITICAL AND ECONOMIC INTERESTS OF ALL THE NATIONS AND PEOPLES OF THE WORLD AS  
13 1  
ELL AS IN THE INTERESTS OF ENSURING THEIR GENUINE SECURITY AND PEACEFUL FUTURE.  
14 1  
THREAT TO INTERNATIONAL PEACE AND SECURITY AND EVEN TO THE VERY SURVIVAL OF  
16 2

く同じようにすればよい。逆に、名詞だけを出力したければ、“**SELECT LSQ**” という単語連鎖選別命令に続く部分を、

**SELECT LSQ**

10

とすればよい。<sup>6)</sup>

このようにして別々に得られた出力を、既に得られたテキスト中の全単語についての出力の該当の個所に挿入しておけばよい。

上掲の単語連鎖選別命令 (**SELECT**) は、テキストの特定の部分についてのコンコーダンスやKWIC索引が必要な場合にも用いることができる。例えばテキストの最初の 500 行についてのみ処理したければ、**CONCORDANCE** 命令あるいは**KWIC INDEX** 命令の後に、

**SELECT LSQ**

1 TO 500

×END

とすればよい。

これに対してテキスト中の全単語についての出力を得る得ないは別にして、テキスト中の特定の語彙 (のグループ)、あるいは語形、接辞等についてのコンコーダンスが有用な場合も、研究目的によっては、ありえよう。このような場合には、**EXCLUDE** か **INCLUDE** で始まる単語選別命令を用いればよい。例えば、従属接続詞についてだけの出力が得たければ、

11LEX

KWIC

INCLUDE

AS

IF

THOUGH

...

×END

[ HALFSIZE ]

などとすればよい。<sup>7)</sup>

単語選別命令の有用性は、特定の接辞を有する語についてのみの出力を得たい時に発揮される。接頭<sup>8)</sup>

辞（あるいは前方一致）の場合は，**C̄ONC̄ORDANCE** 命令にしろ，**KWIC INDEX**命令にしろ，出力はアルファベット順に並べられているからさして問題はないが，接中辞（あるいは中間一致），接尾辞（あるいは後方一致）をアルファベット順リストから探し出すのは必ずしも容易ではない。

まず中間一致の例を挙げる。“ARM”という文字列を含む単語についての出力が欲しい時，**LEX**文は次のようになる。

```
// LEX
KWIC
INCLUDE
SUBSTRING = INFIX
ARM
×END
[ HALFSIZE ]
```

<sup>9)</sup>  
**SUBSTRING** 明細指示で **INFIX** を指定することにより，“ARM”という文字列と中間一致する，即ち“ARM”という文字列を含む単語すべてについての**KWIC**索引が得られる。この出力例を次の図6に示す。

図6 中間一致の例（ARM）

ARMAMENTS ( 1)			
TH A REDUCION IN THE PRESENT LEVEL OF ARMAMENTS.	10	1	
ARMS ( 2)			
ARMS. ADMITTEDLY, THEIR SURVIVAL HAS, IN	4	1	
HANKIND. THE NUCLEAR AND CONVENTIONAL ARMS BUILD-UP THREATENS TO STALL THE	17	2	
ARMS-RACE ( 3)			
THE ENDING OF THE ARMS-RACE AND THE ACHIEVEMENT OF REAL DIS	11	1	
ITS AVENUES ARE CLOSED, THE CONTINUED ARMS-RACE MEANS A GROWING	15	2	
VE TO THE EFFORTS OF STATES TO END THE ARMS-RACE,	23	3	
DISARMAMENT ( 2)			
AND TO SEEK SECURITY IN DISARMAMENT, THAT IS TO SAY, THROUGH A GR	9	1	
ARMS-RACE AND THE ACHIEVEMENT OF REAL DISARMAMENT ARE TASKS OF	11	1	

上の**LEX**文で，**SUBSTRING**明細指示の **INFIX** を**PREFIX**とすれば，“ARM”と前方一致する，即ち，“ARM”という文字列で始まる，単語についての出力となり，**SUFFIX**とすれば，ARM と後方一致する，つまり，“ARM”で終る，単語についての出力となる。次の図7に示すのが，**SUFFIX**とした場合の例で，“ED”で終る単語すべてについての**KWIC**索引の一部である。但し，“BED”，“HEED”



10)  
等は**REJECT**を使って予め除いてある。

#### 図7 後方一致の例(ED)

RELEVED ( 1 )	THAT IN ANOTHER HE SAL BE RELEVED.	4182
ROSTED ( 1 )	FOR ALE AND BREED, AND ROSTED HEM A GOOS,	4137
SAKKED ( 1 )	AND WHAN THE MELE IS SAKKED AND YBOUNDE,	4070
SMYLED ( 1 )	THIS MILLERE SMYLED OF HIR NYCETEE,	4046
SPED ( 1 )	HE AUNTRED HYM, AND HAS HIS NEDES SPED,	4205
SPORNEED ( 1 )	TIL THAT THE MILLERE SPORNEED AT A STOON,	4280
SWYVED ( 2 )	SWYVED THE MILLERES DOGHTER BOLT UPRIGHT,	4266
	HIS WYF IS SWYVED, AND HIS DOGHTER ALS,	4317

図6, 7の例のように, **INCLUDE**と, **SUBSTRING** を用いて単語の選別を行なう時には, 通常, 所謂ノイズ, 即ち不要な単語(上例では, “BED” など)が混入する。レイアウトを問題にしなければ, 実用という目的のためにはさほど支障はないが, 不要な単語を除いた索引だけが特に必要であれば, <sup>11)</sup>**SELECT**や **REJECT** を用いてそのような単語を用いておけばよい。

#### 付 記

以上の説明で, 所謂コンコーダンスの作成が決して難事ではないことを理解していただけたと思う。  
**LEX** は現在開発の途中であり, 不備も多いが, 御意見, 関心のある読者は筆者に一報されたい。

#### 註

- (1) 広島大学計算センターで**LEX**を使う場合。他機種, 他機関の場合は, これに相当する指令(コマンド)が必要である。
- (2) テキストが長い場合には, 一度にすべての単語について行なうことは諸種の制約により不可能と言ってよい。このような場合, データ選別命令を使って何回かに分けて出力することになる。例えば, アルファベット順に分けて行なうとすれば,

// LEX

C̄ONC̄OR

INCLUDE

SUBSTRING = PREFIX

A

×END

というような命令になろう。データ選別命令については次節を参照。

( 3 ) KWIC索引については、例えば文献( 4 )の3.8などを参照。

( 4 ) テキストが長い場合については、上の註( 3 )を参照。

( 5 ) この作業はテキスト中のすべての単語について行なう作業( 本文2節 )とは、別の作業である。

( 6 ) 註5参照。

( 7 ) 実際には、“ as ”，“ that ” などについては、他の品詞の場合と区別するためには、同綴異語について述べたように単語連鎖選別命令( SELECT )を使う必要があろう。

( 8 ) 正常には、特定の文字列で始まる、終る、あるいは特定の文字列を含む、と言うべきである。

( 9 ) LEXで言う中間一致は、前方一致、後方一致、完全一致すべてを含む。

( 10 ) SELECTと逆の機能を有する単語連鎖選別命令。

( 11 ) この場合、LEXに備わったFREQ命令、PRINT WORD FILE命令、INDEX命令などを用いて、条件に合う単語にどのようなものがあり、どの行に出現するかを、まず調べておく必要がある。

## 文 献

( 1 ) 松尾雅嗣( 1979 ) 「テキスト語彙処理プログラムLEXについて( I )」, 「平和科学研究通信」Vol. 3, No. 2, 広島大学平和科学研究センター

( 2 ) 松尾雅嗣( 1980 ) 「テキスト語彙処理プログラムLEXについて( II )」, 「平和科学研究通信」Vol. 3, No. 3, 広島大学平和科学研究センター

( 3 ) 松尾雅嗣( 1979 ) 「テキスト語彙処理プログラムLEXの開発について：概要と論理」 「広島平和科学」2, 広島大学平和科学研究センター

( 4 ) 橋本昌幸( 1971 ) 情報検索のABC, 日本放送出版協会。