テキスト・データベース「トーマス・マン・ファイル」の完成と再編成について

樋口，忠治
九州大学教養部

篠原，武
九州工業大学情報工学部

https://doi.org/10.15017/1468165

出版情報：九州大学大型計算機センター広報．20（6），pp.582-596，1987-11-25．九州大学大型計算機センター
バージョン：
権利関係：

テキスト・データベース「トーマス・マン・ファイル」の完成と再編成について

樋口 忠治*, 篠原 武**

1. はじめに

テキストデータベース管理システムSIGMA[1]によって，Thomas Mann 全集が「テキスト・データベース」としては世界で初めてオンラインで公開されたのは1983年のことであった[2]. 当初は，この全集のうち約半分程度（注①）のテキスト・データの量でスタートしたのであるが，その後，テキスト・データは順次追加され[4]，昨年(1986)の7月をもって全13巻の全集すべてのテキストのデータ化が終了し，それと同時に，SIGMAシステムに格納された.

このたび，SIGMAシステムの改訂[5]を機会に，テキスト・データの再編成を行うことにした. これまでは，システムの効率などの理由から，テキスト・データは全集本の約100～150頁(200～300キロバイト）程度に相当する量に分割して，合計108のファイルによって編成していた. 今回のシステムの改訂により，数メガバイト程度の大きさのファイルでも高速に検索処理が行えるようになったので，原則としてファイルは作品単位にまとめ， また，第8巻から第13巻までのように多数のテキストを集めた巻においては，各巻を単位としてひとつのファイルにまとめた. ただし，第4巻と第5巻は2巻にまたがる1作品であるが，これは2つのファイルに分割している. このような再編成によって，全体は15本の大きなファイルによって構成されることになった.

テキスト・データベースによる検索が可能になったことにより，従来の言語科学では全く手がつけられなかった研究分野が生れた[3]. すなわち，言語研究の分野においてテキスト・データベースは数量化の道を拓いたのである. 研究者の永年の経験と知識に頼っていたこの分野に，だれでも簡単に反復して同じ事実を追体験することができる手段をもたらしたのである.

また，短時間に検索ができるようになったということは，単に時間の問題に留まらず，全体の展望が可能になったことを意味する. なぜならば，個々の問題点を検証するための事実(用例)を拾いあげていくのに何ヵ月も何年も要するのでは，だれもそのような作業をしたがらず，結局は安易な方法を選んでしまうからである（注②）. そのため，新たな事実の発見というようなことは起こりにくく，結局，研究分野全体のエネルギーが歪んだ方向に進みかねない恐れさえある. テキスト・データベースは時間をほとんど無視しうる程度に短縮してくれるのであるから，任意の語や句といった言語表現の用例のすべてを即座に取り出すことが可能であり，したがって，全体的な実像を「一望のもとに」示すことができる. このような手段はかつて言語研究の歴史が手にしたことのなかったものである.

本稿では， SIGMAシステムの改訂にともなって変更されたファイルの利用法を中心に解説を行う. SIGMAシステムの使用法については「トーマス・マン・ファイル」を利用するために必要であると思われる最低限のものを使用例を用いて説明するに留めたので，その詳細については本広報別稿[6]を参照されたい.

昭和62年9月30日受理

*：九州大学教養部

**：九州工業大学情報工学部

2．ファイルの構成

Thomas Mann の著作は，S. Fischer書店より全13巻の全集として出版されている．文学作品は第1巻から第8巻までで，そのうち第8巻は短編のみを集めたものである．第9巻から第12巻までは評論および講演等からなり，全体として“Reden und Aufsätze”という表題がついている．最後の第13巻は補遺であり，雑多な文が入っている．これらに共通なのは公表された文章であることで私的な個人宛の書簡などは含まれていない．「トーマス・マン・ファイル」にはこの全集の全テキストが収められている．

2.1 ファイル一覧

テキスト・データのすべてはSIGMAシステムのファイルを共有するために設けられたSIGMA領域と呼ばれる場所に格納されている．新らしく編成されたファイルの一覧を例1に示す．この一覧は，SIGMAシステムのDDIRECTORYコマンドで表示したものである．OPTIONに“A”を指定しているので，ファイル名とその属性（長さ＝文字数，作成日付時刻など）のほか，コメントが表示されている．作品を納めたファイルには，その出典がわかるように，巻番号，頁番号の始めと終わり，作品名をコメントとして付加している．この一覧から，たとえば，“Die Buddenbrooks”（第1巻の9頁から759頁）は“MANN.BB”という名前のファイルに格納されており，1,737,293文字の大きさのテキストであることがわかる．このように，「トーマス・マン・ファイル」の作品本体のテキストはファイル名の先頭にすべて“MANN”を付した．また，ファイル“MANN.BEISPIEL”には，テスト用サンプルとして，短編の“Tonio Kröger”が納められている．

2.2 ファイルの形式

言語研究のための用例検索の単位は主として文章であり，検索された文章の原作品中での出現位置を知ることも重要な目的の一つである．SIGMAシステムでは検索の単位（＝レコード）は，検索時に指定された記号列によってはさまれた部分として仮想的に設定されるものである．この記号列をレコード区切り語（record delimiter）と呼ぶ．

「トーマス・マン・ファイル」では“#”を文章の区切りとして挿入しているので，検索時にはレコード区切り語に“#”を指定すればよい．さらに，文章の出現位置を表わす7桁の数字列を各文の先頭に付加した．つまり，ファイルの形式は，

…#VVPPPLL 文章 #VVPPPLL 文章 #VVPPPLL 文章 #VVPPPLL …

であり，VVは巻，PPPはページ，LLは行を表わした数字列である．また，パラグラフの区切りとして，“@”を挿入しているので，レコード区切り語の設定を変えれば，パラグラフ単位の検索を行うことも可能である．

ドイツ語文章を計算機に格納する場合には，いくつかの問題が生じるが，「トーマス・マン・ファイル」では，次のような形式で文章を格納することにした．

1) 文章は，“.”（ピリオド）および，“?”（疑問符），“!”（感嘆符），“;”（セミコロン）で区切られるものとして，機械的に処理する．
2) 単語の前後に必ず空白を入れる．
3) アルファベットは，すべて大文字を用いる．
4) 名詞の先頭は本来大文字表記すべきものであるから，その直前に“(”記号を置く．
5) 変母音ä, ö, üはそれぞれ=A, =O, =Uとし，それらが大文字である場合は(=A, (=O, (=Uとする．

（例１）　ファイルの一覧表示とファイルの形式

下線部分は利用者による入力を表す（以下の例も同じ）．

```
DO:DDIRECTORY
FILE:=S. A70152B.MANN.*'
PASSNUMBER:=↙
OPTION:=A
 FILENAME                                        ID ALIAS      SIZE     DATE  TIME
MANN.BB                                          4A     0   1737293 87:09:11 10:28
  BD. I     (  9 - 759)  DIE ⟨BUDDENBROOKS
MANN.BEISPIEL                                 11565     0    160243 87:09:11 11:54
  SAMPLE TEXT FROM ⟨TONIO ⟨KR=OGER
MANN.DF                                          80D     0   1567633 87:09:11 10:30
  BD. VI    (  9 - 676)  ⟨DOKTOR ⟨FAUSTUS
MANN.ERZ                                        148Y     0   2529088 87:09:11 10:33
  BD. VIII  (  9 - 957)  ⟨ERZ=AHLUNGEN
MANN.EW                                         259R     0    598966 87:09:11 10:34
  BD. VII   (  9 - 260)  DER ⟨ERW=AHLTE
MANN.FK                                         285Z     0    939792 87:09:11 10:35
  BD. VII   (265 - 661)  ⟨BEKENNTNISSE DES ⟨HOCHSTAPLERS ⟨FELIX ⟨KRULL
MANN.JS1                                        3265     0   2058062 87:09:11 10:37
  BD. IV    (  9 - 903)  ⟨JOSEPH UND SEINE ⟨BR=UDER (1)
MANN.JS2                                        417D     0   2108189 87:09:11 10:39
  BD. V     (913 -1822)  ⟨JOSEPH UND SEINE ⟨BR=UDER (2)
MANN.KH                                         5090     0    841529 87:09:11 10:40
  BD. II    (  9 - 363)  ⟨K=ONIGLICHE ⟨HOHEIT
MANN.LT                                         546K     0    950185 87:09:11 10:41
  BD. II    (369 - 765)  ⟨LOTTE IN ⟨WEIMAR
MANN.NT                                       10840     0   1647910 87:09:11 11:28
  BD. XIII  (  9 - 872)  ⟨NACHTR=AGE
MANN.RA1                                        690Y     0   2119454 87:09:11 11:23
  BD. IX    (  9 - 951)  ⟨REDEN UND ⟨AUFS=ATZE (1)
MANN.RA2                                        783T     0   2083417 87:09:11 11:24
  BD. X     (  9 - 937)  ⟨REDEN UND ⟨AUFS=ATZE (2)
MANN.RA3                                        8740     0   2567461 87:09:11 11:25
  BD. XI    (  9 -1150)  ⟨REDEN UND ⟨AUFS=ATZE (3)
MANN.RA4                                        987I     0   2227183 87:09:11 11:27
  BD. XII   (  9 - 978)  ⟨REDEN UND ⟨AUFS=ATZE (4)
MANN.ZB                                         5874     0   2347239 87:09:11 10:44
  BD. III   (  9 - 994)  DER ⟨ZAUBERBERG

        TOTAL =  16          PREFIX = A70152B
10777 SECTOR(S) AVAILABLE
DO:LIST S.'A70152B.MANN.ZB'
#@
#0300907 DIE ⟨GESCHICHTE ⟨HANS ⟨CASTORPS , DIE WIR ERZ=AHLEN WOLLEN , - NICHT
UM SEINETWILLEN  ( DENN DER ⟨LESER WIRD EINEN EINFACHEN , WENN AUCH ANSPRECHEN
DEN JUNGEN ⟨MENSCHEN IN IHM KENNENLERNEN  ) , SONDERN UM DER ⟨GESCHICHTE WILLE
N , DIE UNS IN HOHEM ⟨GRADE ERZ=AHLENSWERT SCHEINT ( WOBEI ZU ⟨HANS ⟨CASTORPS
⟨GUNSTEN DENN DOCH ERINNERT WERDEN SOLLTE , DA$ ES SEINE ⟨GESCHICHTE IST , UND
 DA$ NICHT JEDEM JEDE ⟨GESCHICHTE PASSIERT  ) : DIESE ⟨GESCHICHTE IST SEHR LAN
GE HER , SIE IST SOZUSAGEN SCHON GANZ MIT HISTORISCHEM ⟨EDELROST =UBERZOGEN UN
D UNBEDINGT IN DER ⟨ZEITFORM DER TIEFSTEN ⟨VERGANGENHEIT VORZUTRAGEN .
#@
#0300917 DAS W=ARE KEIN ⟨NACHTEIL F=UR EINE ⟨GESCHICHTE , SONDERN EHER EIN ⟨VO
RTEIL ;
#0300918 DENN ⟨GESCHICHTEN M=USSEN VERGANGEN SEIN , UND JE VERGANGENER , K=ONN
TE MAN SAGEN , DESTO BESSER F=UR SIE IN IHRER ⟨EIGENSCHAFT ALS ⟨GESCHICHTEN UN
D F=UR DEN ⟨ERZ=AHLER , DEN RAUNENDEN ⟨BESCHW=ORER DES ⟨IMPERFEKTS .
#0300921 ES STEHT JEDOCH SO MIT IHR , WIE ES HEUTE AUCH MIT DEN ⟨MENSCHEN UND
UNTER DIESEN NICHT ZUM WENIGSTEN MIT DEN ⟨GESCHICHTENERZ=AHLERN STEHT : SIE IS
T VIEL =ALTER ALS IHRE ⟨JAHRE , IHRE ⟨BETAGTHEIT IST NICHT NACH ⟨TAGEN , DAS ⟨
ALTER , DAS AUF IHR LIEGT , NICHT NACH ⟨SONNENUML=AUFEN ZU BERECHNEN ;
#0300926 MIT EINEM ⟨WORTE : SIE VERDANKT DEN ⟨GRAD IHRES ⟨VERGANGENSEINS NICHT
 EIGENTLICH DER ⟨ZEIT , -  EINE ⟨AUSSAGE , WOMIT AUF DIE ⟨FRAGW=URDIGKEIT UND
EIGENT=UMLICHE ⟨ZWIENATUR DIESES GEHEIMNISVOLLEN ⟨ELEMENTES IM ⟨VORBEIGEHEN AN
GESPIELT UND HINGEWIESEN SEI .
#@
......
```

6) 文頭の大文字は，それが名詞である場合を除いて，“〈”記号を付けない．
7) B（エスツェット）はSSではなく通貨記号$で表わす．
8) 引用符号“〈”および“〉”はともに“〉”で表わす．
9) 分離動詞の分離した前つづりはその前に“+”記号を付ける（前置詞と同型のもののみ）．
10)フランス語などドイツ語以外の言語の字飾りは省く．
たとえば，例１で表示しているテキストが“Der Zauberberg”の冒頭部分である．

２．３　ファイル名の指定

ファイル名は2.1に示した通りであるが，利用者が検索対象のファイルなどを指定するために入力するファイル名は，それがSIGMA領域のファイルである場合には，予めプレフィックスの指定をしてない限り，次に示すように「完全名」を与えなければならない．

（例２）　完全名による指定

```
FILE:=S.'A70152B.MANN.ERZ'
```

ファイル名の先頭の“S.”は，それがSIGMA領域のものであることを表わしており，SIGMA領域にあるファイルを指定するときには必須である．次に続く文字列はシングルクォート’で囲まれているが，これは完全名による指定であることを示している．“A70152B.”は，「トーマス・マン・ファイル」が格納されているSIGMA領域の所在場所（＝データベース管理者の利用者ID）を表わしている．ここまでの部分は「トーマス・マン・ファイル」を利用する場合には共通である．それ以降の，“MANN.ERZ”が領域内でのファイル名を表わしている．作品のテキストを納めたファイルの名前はすべて先頭を“MANN.”で統一した．

同一のデータベースを利用する場合，SIGMA領域の所在場所を表わす利用者IDは固定されるので，あらかじめそれをシステム定数のプレフィックスに設定しておくと便利である．プレフィックスの設定は，PROFILEコマンドを用いて，次のように行う．

（例３）　プレフィックスの設定

```
DO:PROFILE
PROF:PREFIX
MEMO OR SIGMA? (M OR S)S
PREFIX :=A70152B
PROF:SAVE
SYSTEM CONSTANTS SAVED
PROF:END
DO:_
```

SAVEサブコマンドは，設定したシステム定数を保存するためのもので，次回のシステム利用時には，そこで保存されたシステム定数が用いられる．SAVEサブコマンドを用いて保存をしなかった場合には，そのセション内でのみ有効な一時的設定となる．「トーマス・マン・ファイル」を主として使用する利用者は，上の例のようにプレフィックスを設定し，それを保存しておくとよいであろう．プレフィックスを用いてファイル名の指定をする場合は，次に示すように「部分名」を与えればよい．

（例４）　プレフィックスを用いた部分名による指定

```
FILE:=S.MANN.ERZ
```

部分名は，“S.” に領域内でのファイル名を続けたものである． この例では，プレフィックスが “A70152B” に設定されているとすれば， 上の完全名で指定したファイルと同じものを指定していることになる．

各自のMEMO領域のファイルなど，SIGMA領域以外のファイルの場合は， 先頭の “S.” を指定せず， たとえば，MEMO区域のファイル “ABC” ならば，次のように，単にその名称のみを指定すればよい．

（例５） MEMOファイルの指定

```
FILE:=ABC
```

その他，ファイル名の指定の詳細については，SIGMAシステムの解説を参照されたい．

３．使用法

「トーマス・マン・ファイル」の使用は，文章の検索が主な目的であると考えられるので，ここでは，検索を行うSEARCHコマンドとその結果を表示したり，再ファイル化を行うREFILEコマンドを中心に，具体的に使用例を挙げながら説明する．

SEARCHコマンドは，テキスト・データを何ら加工することなくそのままの形で前から後ろに一読する間にすべての検索処理を行う逐字サーチの方法を用いている．SEARCHは逐字サーチの方法を有効に活用し，複数の検索キーワードを用いた複数の質問を同時に処理できる特徴をもっている．これにより，通常では何度もファイルを読まなければならないような，動詞の全ての変化形の用例を求めるといった処理を極めて高速に行うことができるのである．

なお，実際のSEARCHコマンドの機能の詳細は文献[6]に示されているので，ここでは読者の理解を妨げぬよう最小限必要な事項の説明に留めておく．

以下の使用例において，下線を施した部分が利用者による入力であり，利用者による入力が復改のみである場合には，見やすさのために↙で示した．また，左の番号は説明に対応するものである．

３．１ 簡単な使用例（TSSセションの開始から終了まで）

```
     JECT005 SYSTEM READY
(1)  LOGON TSS UserId
     + PASSWORD ?
(2)  Password
     KDS70001I UserId   LAST ACCESS AT hh:mm:ss ON yy.ddd
     KEQ56455I UserId  LOGON IN PROGRESS AT hh:mm:ss ON  month  dd, year
     JOB NO = TSUnnnn CN(01)
     KEQ56951I NO BROADCAST MESSAGES
     READY
(3)  SIGMA
(4)  THE CREATION OF SPACE IS STARTED.
     THE NUMBER OF BLOCKS MUST BE FROM 10  TO 4776 .
     NUMBER OF BLOCKS:=50

     NUMBER OF BLOCKS      =   50
     NUMBER OF WORK   FILES=   20
     NUMBER OF LOG    FILES=   20
```

```
     NUMBER OF INDEX  FILES=   20
     NUMBER OF BACKUP FILES=  900

     MASTER KEY:=1111

         1144000 BYTES ARE AVAILABLE.

     THE CREATION PROCESS IS COMPLETED.
(5)  SIGMA)PROF
     PROF:P
     MEMO OR SIGMA? (M OR S)S
     PREFIX :=A70152B
     PROF:END
(6)  DO:DDIR
     FILE:=S.MANN.*
     PASSNUMBER:=↙
     OPTION:=A
      FILENAME                                    ID ALIAS      SIZE     DATE  TIME
     MANN.BB                                      4A    0    1737293 87:09:11 10:28
       BD. I     (  9 - 759)  DIE (BUDDENBROOKS
     ......（例１）のファイル一覧と同じものが表示される......
     MANN.ZB                                    5874    0    2347239 87:09:11 10:44
       BD. III   (  9 - 994)  DER (ZAUBERBERG

              TOTAL =  17          PREFIX = A70152B
     10781 SECTOR(S) AVAILABLE
(7)  DO:SEARCH
          RECORD DELIMITERS
(8)  DO1:=#
     DO2:=↙
          ITEM DELIMITERS
(9)  DO2:=↙
          KEYWORDS
(10) A01:= WENN ... AUCH
     A02:= WIE ... AUCH
     A03:=↙
          LOGICAL FORMULAS
(11) Z01:=A1
     Z02:=A2
     Z03:=↙
(12) REPORT (Y/N)?↙
(13) FILE:=S.MANN.BEISPIEL
        RETRIEVED TEXTS
     QUESTION 01 (Z01) =          3          3
     QUESTION 02 (Z02) =          2          2
     TOTAL             =          5          5
     CPU (SEC/1000)    =         62         62
     FILE:=↙
(14) DO:REFILE
(15) REPORT (Y/N)?N
(16) QUESTION:=1
(17) NEW RECORD DELIMITER:=#
(18) NUMBERING (N/Y)?↙
```

```
(19) OUTPUT-FILE:=W
     QUESTION:=↙
(20) DO:LOOK
     #0828624 ... NEIN , NEIN , SEIN ⟨PLATZ WAR DENNOCH HIER , WO ER SICH IN ⟨INGE'
     S ⟨N=AHE WU$TE , WENN ER AUCH NUR EINSAM VON FERNE STAND UND VERSUCHTE , IN DE
     M ⟨SUMMEN , ⟨KLIRREN UND ⟨LACHEN DORT DRINNEN IHRE ⟨STIMME ZU UNTERSCHEIDEN ,
     IN WELCHER ES KLANG VON WARMEM ⟨LEBEN .
     #0828708 HATTE ETWA NICHT K=URZLICH EINE ⟨ZEITSCHRIFT EIN ⟨GEDICHT VON IHM ANG
     ENOMMEN , WENN SIE DANN AUCH WIEDER EINGEGANGEN WAR , BEVOR DAS ⟨GEDICHT HATTE
      ERSCHEINEN K=ONNEN ?
     #0830809 ZUWEILEN IN DIESEN DREIZEHN ⟨JAHREN , WENN SEIN ⟨MAGEN VERDORBEN GEWE
     SEN WAR , HATTE IHM GETR=AUMT , DA$ ER WIEDER DAHEIM SEI IN DEM ALTEN , HALLEN
     DEN ⟨HAUS AN DER SCHR=AGEN ⟨GASSE , DA$ AUCH SEIN ⟨VATER WIEDER DA SEI UND IHN
      HART ANLASSE WEGEN SEINER ENTARTETEN ⟨LEBENSF=UHRUNG , WAS ER JEDESMAL SEHR I
     N DER ⟨ORDNUNG GEFUNDEN HATTE .
     #
(21) DO:PUT WENN.AUCH
(22) DO:DIR
     FILE:=↙
     PASSNUMBER:=↙
     WENN.AUCH

              TOTAL =   1
      1492 SECTOR(S) AVAILABLE
(23) DO:END
(24) SIGMA⟩END
     READY
(25) LOGOFF
     RETURN CODE : 0000
     ......
```

（使用例の説明）

(1) SIGMAシステムはTSSの配下で処理を行う．TSSセションを開設するために，LOGONコマンドを用いている．"UseriD"は利用者番号である．

(2) パスワードを入力している．パスワードの照合が終わると，TSSセションが開設される． メッセージ"READY"は，システムがTSSのコマンドを受け付ける状態であることを表わしている．

(3) SIGMAシステムを起動するために，SIGMAコマンドを投入している．

(4) SIGMAシステムを利用するためには，個人用の作業場としてのMEMO領域が必要である． 初めてシステムを使用するときなど，MEMO領域がない場合には，この使用例のように，領域の設定が行われる．領域の設定に際して，その大きさ（ブロック数）とマスターキー（ファイルや領域の保護に用いる）をたずねてくるので，それらを指定する．領域の設定が終わると，プロンプト（入力促進文字列）"SIGMA⟩"が端末に表示される．MEMO領域が既に存在する場合には，SIGMAコマンドを投入すると直接この状態へ移る．

(5) "SIGMA⟩"が表示されると，SIGMAシステムのコマンドを実行することができる．ここでは，「トーマス・マン・ファイル」を利用するために， 2.3で説明したように，プレフィックスを設定している．

(6) DDIRECTORYコマンドを用いて，SIGMA領域にある“MANN.”で始まる名前をもつファイルの一覧を表示している．その結果表示されるものは，例1で示したものと同じである．

(7) SEARCHコマンドを用いてテキストの検索を開始している．

(8) SEARCHの検索の単位は，レコード区切り語（RECORD DELIMITERS）と呼ばれる文字列ではさまれた部分でレコードと呼ばれ，検索時に仮想的に設定される．レコード区切り語として任意の文字列を用いることができるが，「トーマス・マン・ファイル」では，2.2で述べたように，文と文の間に“#”記号を挿入しているので，通常は“#”をレコード区切り語に用いればよい．

ここで注意しなければならないことは，もし，レコード区切り語を何も指定しない場合には，ファイル全体が一つのレコードと見なされてしまい，検索結果は0か1のいずれかになることである．レコード区切り語は複数設定することもできるが，ここでは，“#”だけでよいので，復改のみを入力して次の項目区切り語の設定へ移っている．

(9) 項目区切り語は検索の意味の単位を表わすためなどに用いられるが，ここでは用いない．

(10) 区切り語の指定を終えると，次に質問に用いるキーワードの登録をしなければならない．キーワードもまた任意の文字列である．したがって，1個の単語“ NACH ”や2個の単語からなるフレーズ“ NACH〈HAUSE ”も正しいキーワードである．また“+”をキーワードとして用いれば，分離動詞の分離した前つづりを含む文を検索することができ，“ +AUF ”を用いれば分離した前つづりのうち“ +AUF ”を含む文のみを求めることもできる．

SEARCHコマンドではキーワードを同時に複数登録することができ，入力された順にA01, A02, …と名前をつける．SEARCHコマンドでは質問に直接キーワードを用いないで，この名前（キーワード変数とよぶ）を任意に組み合わせて作られる論理式を用いて質問を行う．

指定しているキーワードにトリプルドット“...”が含まれる場合は，文字列の出現順序を指定した検索を行うことができる．ここでは，A01は，“ WENN ”の後に“ AUCH ”が続く文字列を表わし，A02は，“ WIE ”の後に“ AUCH ”が続く文字列を表わしている．キーワードの指定が終わったら復改のみを入力する．

(11) キーワードの登録が終わると質問のための論理式（LOGICAL FORMULAS）を入力しなければならない．論理式は，登録済みのキーワード変数を組み合わせて作る．代表的な組み合わせの例が次に示してある．その意味は右にある通りである．

A1.A2	A1かつA2
A1,A2	A1またはA2
^A1	A1でない

これらの論理式が検索の質問となるのである．SEARCHコマンドでは，このようにして複数のキーワードを組み合わせた複数の質問を指定することができるのである．これ以外の論理式の指定の仕方は文献[6]に詳しい説明があるので参照されたい．ここで注目してほしいことは，SEARCHの検索処理時間は質問の複雑さにはほとんど無関係であるということである．正確にいえば，処理時間は，ファイルを読み込むために必要な，検索するファイルの長さ（文字数）に比例する時間と，検索結果を符号化した形でファイルに出力するために必要な，検索されるレコードの個数に比例する時間の和によって決まるのであり，通常は後者の時間は前者に比べて短いからである．したがって，SEARCHの機能を有効に利用するためには，あらかじめ質問事項をよく検討し，必要であれば適当な短いファイルに対して検索を試行したり

して，この同時処理機能を活用することが望まれる．なお，システムの制限値は，キーワードは99個まで，質問のための論理式は32個までである．

(12)検索の途中経過の状況（質問にヒットしたレコード数や要したCPU時間）のレポートを表示するかどうかの指示を行う．省略時は“Y”を指定したのと同じでレポートが表示される．

(13)論理式の設定で検索の準備がすべて終わる．最後に検索すべきファイル名を入力しなければならない．入力の形式は2.3で述べた通りである．指定されたファイルの検索が終了すると，レポートの表示を行う場合には，ヒットしたレコード数，検索に要したCPU時間（単位はミリ秒）が表示され，さらに検索すべきファイルがあるかどうかをたずねてくる．ここで，ファイル名を入力せずに単に復改キーを押せば，SEARCHコマンドの処理を終了する．

(14)SEARCHコマンドの検索結果を表示したり，再ファイル化を行うためには，REFILEコマンドを用いる．SEARCHの検索結果は，MEMO領域にある索引区域とよばれる場所に，レコードの位置情報と質問番号の関係のみを格納した形で保存される．検索結果の表示は，この索引ファイルから実際のレコードを復元して行われる．

(15)索引ファイルに書き込まれている検索結果のレポート（ヒットしたレコード数）の表示を行うかどうかを指定する．ここでは，検索を行った直後で確認の必要がないので，“N”を指定して表示を行っていない．

(16)どの質問に対する結果を処理するかを質問の番号で指定する．

(17)検索されたレコードを表示したり，再ファイル化したりする際に，レコード間を区切る文字列を指定する．ここでは，もとと同じもの“#”を指定しているが，別の文字列を指定することもできる．

(18)レコードを番号付けするかどうかを指定する．省略時は番号付けを行わない．

(19)出力するファイルを指定する．ここでは，作業ファイル“W”を指定している．特殊ファイル“D”を指定すると，結果は端末ディスプレイに表示される．“K”を指定すると，同様に端末ディスプレイに表示されるが，1レコード毎に中断し，続けて表示を行うかどうかを端末にたずねてくる．出力が終わると，次に処理する質問を入力する．同じ質問に対して処理を繰り返してもよい．ここで，復改のみを入力するとREFILEコマンドを終了する．

(20)LOOKコマンドを用いて，作業区域のトップにあるファイルの内容を表示している．

(21)PUTコマンドを用いて，作業区域のトップのファイルに名前“WENN.AUCH”を付けてMEMO区域に保存している．このファイルは通常のテキストファイルであるので，SEARCHコマンドを用いてさらに検索を実行することもできる．

(22)DIRECTORYコマンドでMEMO区域のファイル名一覧を表示している．

(23)ENDコマンドを用いて，“SIGMA)”状態へ戻っている．これで，最初に“SIGMA)”の状態で使用したコマンドから，このENDコマンドまでの会話記録（LOG）がLOG区域のトップに作られる．LOGファイルはそのまま一種のコマンドプロシジャとして用いることができる．利用法については，次の使用例で簡単な説明を示すが，詳細については文献[6]を参照されたい．

(24)ENDコマンドを用いて，SIGMAシステムを終了させている．このように，“DO:”の状態と“SIGMA)”の状態では，ENDコマンドの意味が異なっているので注意しなければならない．

(25)LOGOFFコマンドを用いてTSSセションを終了している．TSSセションを終了するコマンドには，LOGOFFCコマンドもある．LOGOFFCコマンドを用いると，そのTSSセションで使用した計算機資源（CPU時間，接続時間，入出力行数など）と課金情報が出力される．

3.2 LOGファイルの利用（ファイル名の一括入力，文番号の抽出）

```
     DO:SEA
          RECORD DELIMITERS
     D01:=#
     D02:=↙
          ITEM DELIMITERS
     D02:=↙
          KEYWORDS
     A01:= =AHNLICH SEHEN
     A02:= =AHNLICH SIEHT
     A03:= =AHNLICH SAH
     A04:= =AHNLICH GESEHEN
     A05:= SEHEN...=AHNLICH
     A06:= SIEHT...=AHNLICH
     A07:= SAH...=ANHNLICH
     A08:=↙
          LOGICAL FORMULAS
     Z01:=A1,A2,A3,A4
     Z02:=A5,A6,A7
     Z03:=Z1,Z2
     Z04:=↙
     REPORT (Y/N)?↙
(1)  FILE:=.S.FNAME.ROMAN
     FILE:=S.'A70152B.MANN.BB'
        RETRIEVED TEXTS
     QUESTION 01 (Z01) =          1          1
     QUESTION 02 (Z02) =          2          2
     QUESTION 03 (Z03) =          3          3
     TOTAL             =          3          3
     CPU (SEC/1000)    =        760        760
     FILE:=S.'A70152B.MANN.KH'
        RETRIEVED TEXTS
     QUESTION 01 (Z01) =          1          2
     QUESTION 02 (Z02) =          2          4
     QUESTION 03 (Z03) =          2          5
     TOTAL             =          2          5
     CPU (SEC/1000)    =        373       1133
     FILE:=S.'A70152B.MANN.LT'
        RETRIEVED TEXTS
     QUESTION 01 (Z01) =          2          4
     QUESTION 02 (Z02) =          0          4
     ......
     FILE:=S.'A70152B.MANN.ZB'
     ......
     FILE:=S.'A70152B.MANN.JS1'
     ......
     FILE:=S.'A70152B.MANN.JS2'
     ......
     FILE:=S.'A70152B.MANN.DF'
     ......
     FILE:=S.'A70152B.MANN.EW'
     ......
```

```
     FILE:=S.A70152B.MANN.FK
        RETRIEVED TEXTS
     QUESTION 01 (Z01) =          0         11
     QUESTION 02 (Z02) =          2         19
     QUESTION 03 (Z03) =          2         25
     TOTAL             =          2         25
     CPU (SEC/1000)    =        415       5704
     FILE:=.E
     FILE:=↙
     DO:REF
     REPORT (Y/N)?↙
        RETRIEVED TEXTS
     QUESTION 01 (Z01) =         11
     QUESTION 02 (Z02) =         19
     QUESTION 03 (Z03) =         25
     TOTAL             =         25
     QUESTION:=1
     NEW RECORD DELIMITER:=#
     NUMBERING (N/Y)?↙
(2)  OUTPUT-FILE:=K
     QUESTION 01 (Z01) =         11
     #0106002 " SIEH NUR , SIE HAT STUPENDE ZUGENOMMEN ...  " " WOLLT IHR MIR GLAUB
     EN , DA$ SIE ⟨NETTEN =AHNLICH SIEHT ?
     MORE (Y/N)?↙
     #0226312 " JA , SEHEN ⟨SIE , ⟨PRINZ , DAS WAR NUN WIEDER SO RECHT EINE ⟨FRAGE
     , DIE ⟨IHNEN =AHNLICH SIEHT , EINE AUSGEMACHTE ⟨PRINZENFRAGE .
     MORE (Y/N)?↙
     #0256519 OB ER ⟨CHRISTIANEN =AHNLICH SAH ?
(3)  MORE (Y/N)?N
     QUESTION:=2
     NEW RECORD DELIMITER:=#
     NUMBERING (N/Y)?↙
     OUTPUT-FILE:=K
     QUESTION 02 (Z02) =         19
     #0129426 ABER ES WAR ⟨KONSUL ⟨BUDDENBROOK ... ES SAH IHM =AHNLICH .
     MORE (Y/N)?N
     QUESTION:=↙
(4)  DO:.S.LOG.NUM
     DO:ECHO OFF
        RETRIEVED TEXTS
     QUESTION 01 (Z01) =         11
     QUESTION 02 (Z02) =         19
     QUESTION 03 (Z03) =         25
     TOTAL             =         25
(5)  QUESTION:=3
     QUESTION 03 (Z03) =         25
     DO:END
(6)  DO:LOOK
      #0106002 #0129426 #0151308 #0213903 #0226312 #0256519 #0267707 #0303807
      #0329901 #0367414 #0388822 #0421513 #0421716 #0441712 #0442205 #0474814
      #0528106 #0557117 #0557211 #0557211 #0604118 #0618903 #0714821 #0734502
      #0761713
     DO:_
```

（使用例の説明）

(1) ファイル名ファイルの利用. 出力データの量が極めて少ないことが予想される場合など，すべてのファイルを検索する必要が生ずることがある．そうした場合には，いちいちファイル名を投入していくよりも，複数のファイル名をまとめて投入したり，全ファイル名を一括して投入することができると便利である．これを実現する方法に，ファイル名を格納した「ファイル名ファイル」の利用がある．「トーマス・マン・ファイル」では，利用者の便宜のために，いくつかのファイル名ファイルを作成しているので，これらのファイルを利用するとよい．作品のテキストは“MANN.”で始まる名前のファイルに格納しているのに対し，ファイル名ファイルは“FNAME.”で始まる名前のファイルに格納している．この例で用いているファイル“S.FNAME.ROMAN”（完全名は“S.'A70152B.FNAME.ROMAN'”）には，第１巻から第７巻までの長編すべてのファイル名が

```
FILE:=S.'A70152B.MANN.BB'
FILE:=S.'A70152B.MANN.KH'
......
FILE:=S.'A70152B.MANN.FK'
FILE:=.E
```

の形式で書かれている．端末のキーボードから個々のファイル名を入力する代わりに，先頭にピリオド“.”をおいて，それに続けてファイル名を入力すれば，このファイルから連続的にファイル名が読み込まれる．各自で任意のファイル名を組合せたファイル名ファイルを作って利用することもできる．これは，下の(4)と同様に，LOGファイルの実行機能を用いて実現したものである．詳しい点については，文献[6]を参照されたい．

(2) レコード毎の出力. REFILEの出力ファイルに“K”を指定しているので，レコード毎に出力を中断し，続けて表示を行うかどうかを端末に聞いてきている．

(3) “N”を入力して，“K”へのレコード毎の表示を打ち切っている．

(4) 「文番号」のみの抽出. 検索結果の出力量が過大である場合など，一応その文の位置を示す「文番号」だけを得たいという時には，この例で示されているような方法がある．すなわち，検索が終わった時点で，“.S.LOG.NUM”を投入することによって，検索結果のリファイルを行い，その結果のファイルから「文番号」のみを取り出して一覧表を作成するという一連の作業をすべて行い，そのリストが作業区域のトップのファイルに作成されるようになっている．この一連の作業は，ファイル“S.LOG.NUM”（完全名は“S.'A70152B.LOG.NUM'”）に書かれているSIGMAのコマンドを実行して行われている．こうした一連の作業を実行するためのLOGファイルは，ファイル名ファイルと同様に，“LOG.”で始まる名前のファイルに格納しているので，参考にされたい．

(5) 文番号を取り出す質問を指示している．

(6) 取り出された文番号は，作業区域のトップに作られるので，LOOKコマンドによってこれを端末に表示している．

3.3 キーワードの設定について

不変化詞の場合はキーワードの指定に問題は余りないが，動詞，形容詞などの検索の場合にはキーワードの指定を誤ると，探している語を除外したり，逆に不必要なものまで拾ってしまうこ

とになる．したがって，事前の準備を十分に行うべきである．ある動詞の可能な変化形をすべてキーワードとして指定する方がよいか，または，他に同一の綴りがない場合には語幹などをキーワードとして指定するかはケースバイケースで判断することになる．空白，すなわち語と語（記号の場合も同じ）の間にあるブランクは非常に重要な意味をもっており，これを指定するか否かで結果に大きな相違が生じると考えなければならない．

検索の結果，不必要な語形まで拾ってしまった場合には，その結果をリファイルし，このファイルを対象として，再び検索することによって，不必要なものを除外した結果を得ることが出来る．この方法を利用することによって，第１段階では多少の余分なものまで拾ってしまうとしても，第２段階でより厳密なキーワードの指定をするという，いわば結果を見ながら検索方法（キーワードの指定）を考えるというやり方も可能である．

Kann man die Zeit erzählen, diese selbst, als solche, an und für sich ? (0374807)

上記の文がどこに出ているのか知りたいという場合には，この文全体をキーワードとして指定することもできるが，次のように，先頭の数語を指定するだけでも十分であろう．

A01:= KANN MAN DIE ⟨ZEIT ERZ=AHLEN

これと全く同じテキストを含む文が他にそれ程多くあるとは考えられない．このような指定の他に，任意の文字列を表すトリプルドット“...”を用いる方法がある．たとえば，

A02:= KANN MANN ... AN UND F=UR SICH ?

あるいは

A03:= KANN ... ⟨ZEIT ... ALS SOLCHE

などとすれば，このような語の組合せは稀であるから，出力レコードの数は限られてくる．

４．テキスト・データベースの意義

人類が文字を使用するようになって以来，すでに３千年以上を経たが，その歴史の中でも紙と印刷の発明は画期的な出来事であった．そしてJohann Gutenberg(1394?-1468)による活字印刷法の普及によって今日までの印刷文明が築かれてきたのである．そして今世紀半ば以降に発達したコンピューターが文字と印刷の歴史にとって画期的な意味をもっていることは論を待たない．

言語研究の歴史の中では古来，著名な詩人や作家の作品が語法の用例として示されることになっている．なぜなら，その影響力が大きいからであり，単に一個人の言語ではないことが明らかだからである．そうしたテキストの量は厖大なものであり，とうてい記憶にたよることはできない．そこで，索引が作られることがある．しかし，「索引」というのはせいぜい名詞について，あるいは人名や地名といった固有名詞について作成される程度に留っている．なぜなら，「名もない」ごく普通の語，例えばある動詞や接続詞は「文学的」にはほとんど意味がないからである．ところが，言語の研究にとってはそのような「索引」はほとんど役に立たない．全ての語の「ふるまい」が言語研究の対象であり，しかも可能ならば，その「すべての」ふるまいが明らかになっていることが望ましい．

トーマス・マン・ファイルは「テキスト・データベース」として初めて誕生したものである．研究の方法論自体もまだ十分明らかになっているわけではない．それにも増して，このテキスト・データベースはまだ一作家の全集にすぎない．言語の研究にとっては大海の一滴にすぎないのである．今後，このようなテキスト・データの作成が進み，従来の言語研究で取り扱われてきた範囲に匹敵する程の量のテキスト・データが備わるようになるにはまだかなりの時間が必要であ

ろう．

他方，現存のテキスト・データベースから得られる情報は，わが国内では大学間ネットワークの利用により容易に入手できる状態にある．しかし，「文科系」の分野では研究費は書物や文献を購入するためのものであるという永年の習慣が先入観として身についており，加えて，研究費自体，端末装置などを購入するには余りにも少なすぎるのが現状である．

また，テキスト・データベースは研究者が自分で直接オンラインで検索をして初めてその効力を発揮するものであるから，「トーマス・マン・ファイル」の場合には，日本国内のみならず，外国，特に東西両ドイツから直接アクセスできることが望まれる．しかしながら，現状ではこれを実現することは一利用者の立場では不可能である．この問題を解決する道が何らかの形で切り拓かれることを切に希望する．

5．おわりに

テキスト・データベース「トーマス・マン・ファイル」は少なくとも四つの要素から成り立っている．すなわち，このファイル自体とテキスト・データベース管理システムSIGMA，九州大学大型計算機センターのシステム，大学間ネットワーク（注③）の四つである．これらに加えて，国際通信回線の利用による国際ネットワークが，できるだけ近い将来において，実現されることを期待したい．もし，それが無理であるとしても，ヨーロッパのどこか（例えば，ベルリンの「日独センター」など）に1箇所でもよいから，オンラインでアクセス可能な窓口を設けることができればと考える．「文化の交流」は何も美術品や工芸品の展示だけに限らない．テキスト・データベース「トーマス・マン・ファイル」の検索サービスをヨーロッパに提供することは測り知れない「文化的」、「外交的」な効果を生みだすであろう．その意味を行政関係者が理解されることを期待する．

最後に，今回のファイルの再編成の作業にあたって，九州大学総理工・有川節夫教授をはじめ同研究室の学生達に大変お世話になった．特に，宮原哲浩，川崎洋治，井上仁の三人には，新システムへのファイルの移行で長時間の作業を手伝っていただいた．ここに，深く感謝の意を表します．

注① Ⅰ，Ⅱ，Ⅲ，Ⅵ，Ⅶの各巻及び第Ⅷ巻の主要部分.
注② 既存の辞書類や，研究書の用例から採集すること．したがって，新らたな知見を加えていくことにはならない.
注③ 全国共同利用大型計算機センター「オンライン・データベース利用ガイド」（各センター常備）を参照されたい.

参考文献

1. 有川，篠原，白石，玉越：研究者向き情報システムSIGMAについて，九大大型計算機センター広報，Vol.14，No.4，pp.550-573 (1981).
2. 樋口，篠原：公用データベース トーマス・マン・ファイル／SIGMAの公開，九大大型計算機センター広報，Vol.16，No.4，pp.379-393 (1983).
3. 樋口：テキスト・データベース（SIGMA）－ファイルの取り扱いについて－，九大教養部言語研究会紀要『言語科学』第19号，(1984).
4. 樋口，篠原：公用データベース「トーマス・マン・ファイル」のファイル追加について，九大大型計算機センター広報，Vol.18，No.2，(1985).
5. 有川ほか：テキストデータベース管理システム SIGMAについて，情報処理学会研究報告，Vol.87 No.65 (FI-6-4) (1987).
6. 有川ほか：テキストデータベース管理システムSIGMA第２版について，九大大型計算機センター広報，Vol.20，No.6 (1987).

＊使用上の注意

「トーマス・マン・ファイル」は，西独S.Fischer社の許可を得て（下の協定書参照），専ら言語研究のために提供するものである．このテキスト・データを何らかの方法により加工して，読書用のテキストとして使用することは禁止する．この点には注意を喚起しておきたい.

また機械可読なデータとしてのテキスト・データ自体の“著作権”は著者の樋口にあるので，みだりにコピーを作るなどの方法はとらないようにしていただきたい.

協定書

（概要）

1 トーマス・マンの全作品をドイツ語で電子情報として蓄え，これを科学研究の目的に限り，語いの研究ないしは統計などの言語研究を行なう権利を認める．更に言語研究者および独語独文学研究者に対して大学間ネットワークを通じて無料でデータを提供することも認める.
2 利用者は本協定を順守する義務を負う.
3 データは営利を目的として利用してはならない.
4 データの一部または全部を購読などのために印刷する場合は予め同意を必要とする.
5 データ及び研究成果あるいは利用価値をもつ成果は要求があれば送付しなければならない.
6 本協定の他にドイツ法および著作権法が適用される.
7 判籍はフランクフルト・アム・マインとする.

以上を同意する． １９８２年１１月１０日

S. フィッシャー書店 署名 九州大学教授 樋口忠治 署名