

STAIRS/VS における検索方式と問題点

間下 浩 之 (日本IBM(株)サイエンティフィックセクター)

1. はじめに

数年来コンピュータによる情報検索システムが各方面で開発されてきた。それらの中にIBMのプログラム・パッケージとしてSTAIRS/VS¹⁾(Storage and Information Retrieval System / Virtual Storage)がある。

STAIRS/VSは定型データに限らず、自然語・自然文のような非定型データをもデータベース内に持ち、それらに対する柔軟性に富んだ多彩な検索コマンドを有している。それゆえ多くの論文、報告書、契約書、法律、官報あるいは出版物の要約などを、直接利用しなければならない所で、これは有効な道具となると思われる。本稿では、そのSTAIRS/VSに因りて特にファイル構成、検索方式について説明し、一般の情報検索システムのデータベースを日本語の文章にした場合、どのような問題点が生ずるかを検討する。

2. STAIRS/VS

現在STAIRS/VSには、本来の機能に加えていろいろな付加機能を持ったシステムがある。ここでは、それらのうちから主なものとして、本来の機能としてのSTAIRS/VSとミソラスを導入したSTAIRS/VS-TLS²⁾について紹介することにする。

2.1 STAIRS/VS

STAIRS/VSの機能と、特に辞書であるデータベースについてふれ、簡単なコマンド体系と例を示す。

2.1.1 機能

① オンライン・対話式検索

IBMの情報表示装置3270を使用し、ライトペン等により検索することが可能である。対話サブシステムであるAQUARIUS(A Query And Retrieval Interactive Utility System)のコマンド体系については後述べる。

② 定型データ、非定型データ両方の検索が可能

各項目に構造化が可能なデータ、例えば著者名、題目、出版月日、などは定型データとして取扱う。記述された文章に対しては、それをそのまま入力し、その文章から自動的にキーワードを引き出しインデックスを作成する。検索の時には、著者名や題目に対する向合わせと同時に、キーワードとキーワードの相互関係から文脈に対する向合わせも可能である。

③ 機密保護

情報検索のうえで一番問題となる機密保護が、端末およびデータベース(データベース、文献、文節、文章)の両面からきめこまかに配慮してある。

④ オンライン・データベース作成

大量データの入力に対してはバッチ方式で行なうとしても、データベースの変更、削除に対してはATMS(Advanced Text Management System)のモジュールを使用してオンラインで簡単におこなえる。

2.1.2 データベース作成

データベースを作成する場合、前にもふれたが2種類の作成方法がある。1つはオンライン方式、1つはバッチ方式で、ここではバッチ方式について述べることにする。ユーザーは自分のもっている情報をSTAIRS/VSに入力する際に、システムの入力フォーマットに変換するか、又はそのフォーマットに従って、データを作成しなければならない。フォーマットを図1に示す。IDは、その情報一単位におけるIDを示す。パラグラフとは、データの性格を表すためのものであり、3桁の頭が数字であれば英数字(英数字)でよい。図書館等の3桁の分類と同じと考えて、同じ性格のフィールドは同じ番号をふることに望ましい。テキストの部分はパラグラフ別に定型データ、非定型データが混在して良い。特に非定型データの場合は何十枚、何百枚のカードにまたがって同じパラグラフコードをふり、を入力する。

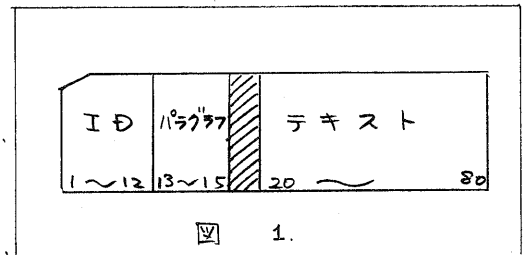


図 1.

このようなデータをデータベース作成プログラムの入力とし、適当なコントロールカードを与えることにより、STAIRS/VSのデータベースが作成される。コントロールカードの主なものとして次の3種類があげられる。

① ストップワード

入力される非定型データ中で、検索対象とならない語をとりのぞく。例えば、助詞、助動詞、冠詞、前置詞・・・である。またこのデータには前一致による消去も可能である。例えばPROGRAM\$3と入力すると、そのあとの数字の桁だけつく語がディクショナリーから消去される。PROGRAMとかPROGRAMMEは消去されPROGRAMMINGは消去されない。

② ストップパラグラフ

指定するパラグラフが、検索対象であるのか、それとも表示だけであるのかを決定する。通常はそのコードを指定するが、2文字とか3文字と入力すると頭の1桁が2の部類に属するパラグラフはすべてディクショナリーから取り除かれる。

③ 定型データの指定

どのパラグラフが、定型データであるのかを指定する。この場合もストップパラグラフと同様にその指定が可能である。又、データの長さや何番目のパラグラフを入力する。

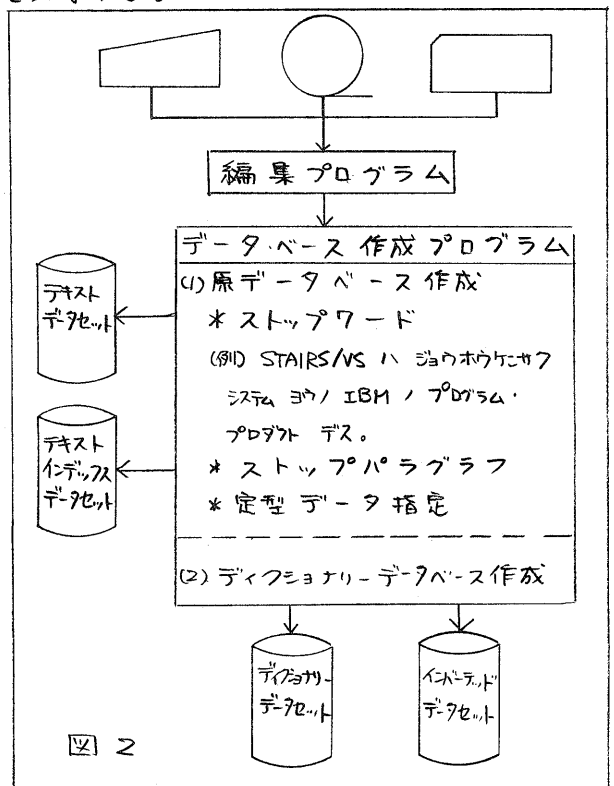
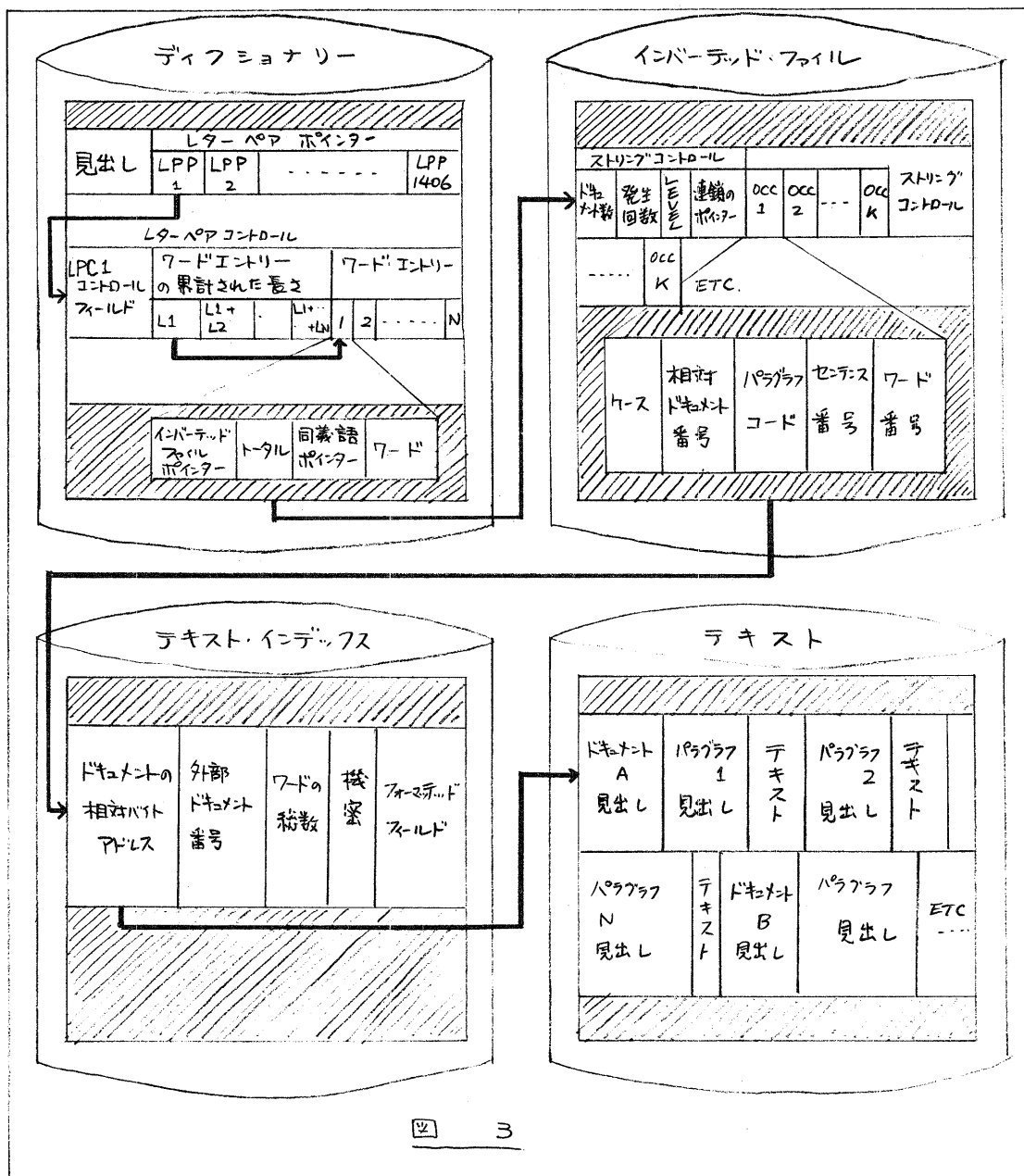


図 2

以上により、データベース作成プログラムはテキストインデックスデータセットとテキストデータセットを作成する。定型データは前者に、非定型データは後者に含まれる。次にデータベース作成プログラムは、入力されたデータをもとに各種情報を自動的にディフショナリーデータセット、インバーテッドデータセットとして作成する。以上の4つのデータセットがSTAIRS/VISにおけるデータベース(辞書)である。

2. 1. 3 データベース

4つのデータセット間の関係を図3に示す。ここではその検索プロセスに従って各々のデータセットについてふれることにする。



①ディクショナリーデータセット

始めに、ある検索ワードが与えられると、その頭の2文字をとって図4に示すLPCマトリックスを調べる。

$D = 4 * [(T(X) - 1) * 38 + T(Y)]$
により何番目のLPC内に該当する検索ワードが含まれているかがわかる。LPC内にはその2文字が同じワードのオカレンスがすべてあり、それらのフィールドには、インバーテッドファイル

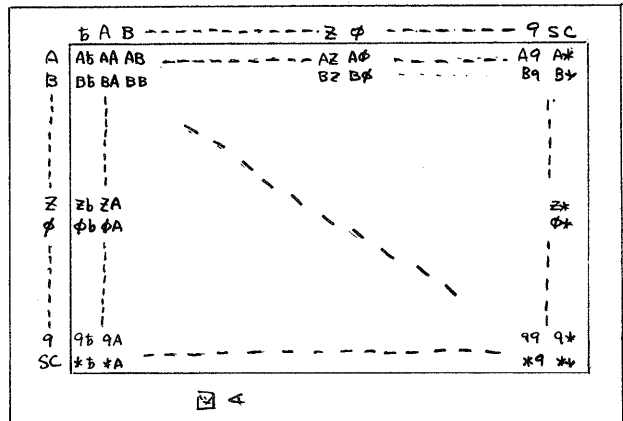


図4

へのポインター、そのワードの現われるドキュメント数、オカレンス数、頭2文字を取り除いた残りがはいつている。同義語ポインターは、データベース生成時にコントロールカードで指定された場合、ワード間に関係がもたれる。

②インバーテッドファイル

ディクショナリーデータセットに含まれるワードそれぞれに対しての内容が含まれている。ディクショナリーから与えられたレコードには、ドキュメント数、オカレンス数、連鎖ポインター等があり、連鎖ポインターから与えられたワードのオカレンスそれぞれには、ドキュメント番号、パラグラフコード、文番号、ワード番号がはいつている。こゝまでで、その検索ワードが、何回発生していて、ドキュメントとしてはいくつあって、それぞれに対して何行目の何番目のワードかを知らることができる。

③テキストインデックスデータセット

ドキュメント番号をキーにして作成されたデータセットである。各レコードには、非定型データのドキュメントのあるレコードの相対バイトアドレス、ドキュメントのID番号、機密レベル、定型データの内容等が含まれている。

④テキストファイル

すべての非定型データがはいつている。可変長のテキストの為、頭にその長さを示すレコードとして、ドキュメントの見出し、パラグラフの見出しがはいつている。

STAIRS / VSの辞書は以上の4つのファイルから構成されていて、AQUARIUSのコマンドによって適当な所でファイルの中をサーチする。

2. 1. 4. オンライン対話機能(AQUARIUS)

大きく分けて3つの機能があり、それぞれにコマンドが用意されている。

①検索機能

・非定型データ検索(..SEARCH)

Bool論理検索手法によって非定型データの検索が可能。

・定型データ検索(..SELECT)

形式化したデータセットの検索手法。

・定型データソート(..SORT)

前に行なわれたSELECTによる検索結果の内容を降順、昇順にソート。

・重みづけ検索手法(..RANK)

指定した適当なアルゴリズムに従ってドキュメントを重要度順に示らる。

②表示印刷機能

- ・データの表示(..BROWSE)
検索したドキュメントの内容を順番に表示。

③補助機能

- ・データベース変更(..CHANGE)
- ・検索質向表示(..DISPLAY)
- ・STAIRSガイド(..HELP)
- ・質向の保存/実行/削除(..SAVE, ..EXEC, ..PURGE)

これらを適当に使用してIBMのマニュアルの検索の例を以下に紹介する。

```
R0201 ENTER ONE OF THE FOLLOWING COMMANDS:
..BROWSE ..SEARCH ..SELECT ..EXEC ..RANK ..HELP ..CHANGE ..OFF
..search
AQUARIUS - SEARCH MODE - BEGIN YOUR QUERY AFTER THE STATEMENT NUMBER
00001program and linguistic$
PROGRAM                                39364 OCCURRENCES      12897 DOCUMENTS
LINGUISTIC$                            9 OCCURRENCES
$                                       3 OCCURRENCES          1 DOCUMENT
$AL                                    1 OCCURRENCE          1 DOCUMENT
$$                                    3 OCCURRENCES          2 DOCUMENTS
$$/LANGUAGES                          2 OCCURRENCES          2 DOCUMENTS
RESULT                                2 OCCURRENCES          2 DOCUMENTS
00002natural adj language
NATURAL                               81 OCCURRENCES          44 DOCUMENTS
LANGUAGE                             5652 OCCURRENCES       2735 DOCUMENTS
RESULT                                9 OCCURRENCES          7 DOCUMENTS
000032 not system
2 NATURAL ADJ LANGUAGE                9 OCCURRENCES          7 DOCUMENTS
SYSTEM                              35602 OCCURRENCES     13790 DOCUMENTS
RESULT                                3 OCCURRENCES          3 DOCUMENTS
00004..browse
AQUARIUS - BROWSE MODE - ENTER PRINT SPECIFICATIONS - OR:
ALL - TOTAL DOCUMENT, A - ALL PARAGRAPHS, B - ALL FIELDS, OR C OR EXPLAIN
explain
I4513 * FOLLOWING FORM. FIELDS ARE AVAILABLE
      SHOWN BY NUMBER, NAME, TYPE, LENGTH :
      1. CLASS      A  5      2. DATE      A  6      3. ORDERNO  A 16
I4514 * FOLLOWING PARAGRAPHS ARE AVAILABLE (SHOWN BY NUMBER, NAME):
      1. TITLE      2. DESCRIBE      3. SECURITY      4. ORDERNO
      5. ABSTRACT   6. INDICES      7. REQUEST
I4515 * PRINT-SPECIFICATIONS ARE A LIST OF FIELD-SPECIFICATIONS, PARAGRAPH
      NAMES AND/OR NUMBERS - SEPARATED BY COMMAS.
I4516 * A FIELD-SPECIFICATION IS: F=( LIST OF FIELD NAMES AND/OR NUMBERS )
R4511 * ENTER YOUR SPECIFICATIONS FOR PRINT SELECTION NOW.
ALL - TOTAL DOCUMENT, A - ALL PARAGRAPHS, B - ALL FIELDS, OR A SPECIFIC LIST.
AQUARIUS - BROWSE MODE - ENTER PRINT SPECIFICATIONS - OR:
ALL - TOTAL DOCUMENT, A - ALL PARAGRAPHS, B - ALL FIELDS, OR C OR EXPLAIN
1,5
BOIS00024820 DOCUMENT=          1 OF          3 NUMBER OF LINES =          25
TITLE      QUERY LANGUAGE ONE (QL/I) USERS MANUAL
           LOS ANGELES SCIENTIFIC CENTER TECHNICAL REPORT
ABSTRACT   QL/I is a problem-oriented language designed for a
           particular class of problems (encountered in medicine as
           specialty or specialties.
```

END OF DOCUMENT

```
..select
AQUARIUS - SELECT MODE - ENTER EXTENT AND CRITERIA AFTER THE STATEMENT
NUMBER OR ENTER EXPLAIN
00005explain
I2516 THE FOLLOWING FIELDS ARE AVAILABLE (SHOWN BY NAME|PARAGRAPH|LENGTH|TYPE):
CLASS      05 A      DATE      06 A      ORDERNO      16 A
00005a11 date w1 720101 760101
I2516 NUMBER OF SELECTED DOCUMENTS = 020369
```

2.2 STAIRS/VS-TLS (Thesaurus and Linguistics integrated System)

STAIRS/VS に新しく追加された機能として大きく以下の3点があげられる。

① シソーラスのサポート

TLS による情報の流れが、図5に示してある。シソーラスを使用した場合はシソーラスエントリーから検索ワードを入力する。よって㉔あるいは㉕のような流れになる。シソーラスを通過することによって指示された範囲のワードが関連づけられ、それが STAIRS/VS の検索ワードとなる。

② 語変化の自動生成機能

検索ワードが最終的に拡張検索ワードに致るまでの過程を簡単に説明する。

1) 入力された検索ワードは、その言語に従って(英語は英語、ドイツ語はドイツ語)その語から考えられる上位集合をとる。

2) それらがディフショナリーにあるかどうかのチェックをする。

3) それらの中で元の検索ワードの正しい変形であるかのチェックをする。これらの段階を通して拡張検索ワードが生成される。これは図5の㉖の流れである。㉕の機能とあわせて㉖のような流れとする。

③ シソーラス作成機能

ユーザーは図6の上部のようき定義をすることによって、自分用のシソーラスの概念を作成することができる。シソーラスジェネレーターはこれらの定義に従って、ユーザーのシソーラスに対する追加、変更、削除等のチェックをし、新しいシソーラスを作成する。

図6は上部の定義に従って、追加と削除を行なった時の流れである。

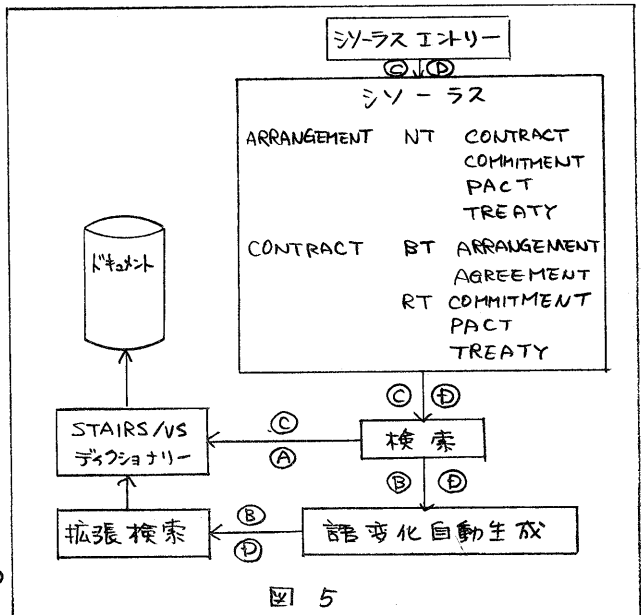


図 5

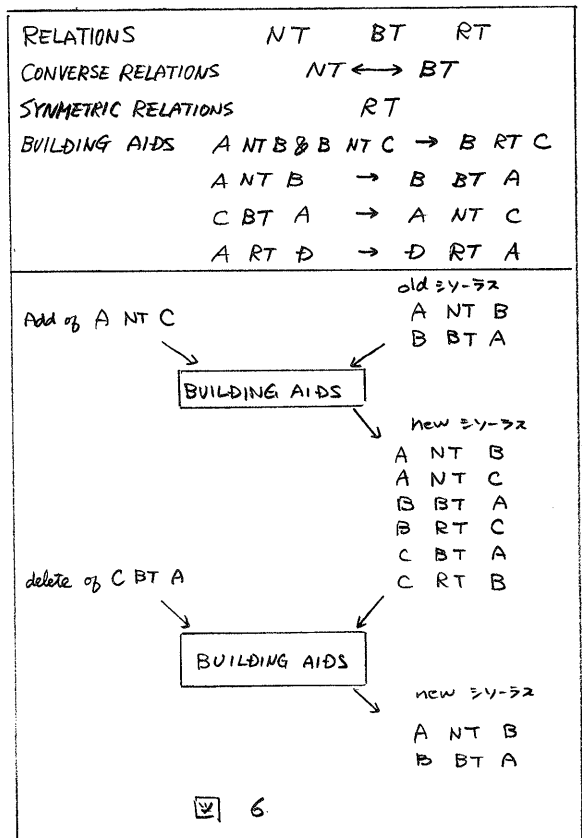


図 6

3. 日本語における問題点

ミソラを使用しない情報検索システムによって日本語における文章を取り扱う場合、どのような問題点が生ずるかを述べる。

3. 1 データベース作成の為の入力

非定型データの日本語と英語における大きな違いは分ち書きにある。自動インデキシングをすることは、自立語と付属語を分離し、自立語内でも重要語だけを取り出すことである。あるいは重要語以外を削除することである。まったく分ち書きされていない文章から重要語を取り出すのは、すべてがカタカナの場合すこし困難であろう。漢字カナ混り文の場合にコンピュータで自動分ち書きをするのは可能性があるかもしれない。入力文をすべての品詞に分けて入力することは人手がかりすぎて実用性に乏しい。適当な漢字ディスプレイがある場合、漢字鍵盤による入力は比較的簡単に思われるが、漢字ディスプレイが無い場合あるいはカナ鍵盤の場合にはここでカナ漢字変換を行なう作業が生ずる。又同音異義語が生じて、人間との対話からそれを解消することにおいても、漢字ディスプレイが無いとその工夫をしなければならぬ。

3. 2 データベースと検索

データベースには入力された文章を漢字カナ混り文そのままに持つが、その文章に対して漢字カナ変換を行ない読み易くも同時にもつかの2つの方法が考えられる。漢字カナ混り入力で検索する場合には、前者で十分である。カナ入力の場合には、後者の方法をとるかあるいは検索文のカナ漢字変換を行ない前者の方法をとるか2種類考えられる。カナ検索文入力の場合には、どちらにしても何万語かを所有する国語辞典をシステムにもっていなければならぬであろう。又、入力文中に合成語(転成語、派生語、複合語)があった場合、それらを分離して辞書を構成するか、合成語独自の辞書を持つかを考えねばならない。検索の時に人によって送りが異なるという点も問題となるだろう。

日本語を取り扱う情報検索システムを設計するには以上の様な問題点をよく考慮した上で行なう必要がある。

4. おわりに

1章2章でSTAIRS/VISとSTAIRS/VIS-TLSの説明をしたが、IBMのプログラムパッケージには他にSTAIRS/DOS/VIS、STAIRS-DL/Iがあげられる。STAIRS-DL/Iは従来のSTAIRSにおけるデータベース、4つのファイル(BDAM)をDL/Iのファイルにすることによって、データベースの維持、保守の容易性の向上、検索結果を別のアプリケーションプログラムで使用可能、等の特長がでている。ユーザーはこれら4つのシステムから自分の要求にあったシステムを使用されたい。3章における日本語の情報検索システムにおける問題は、結局使い易い漢字ディスプレイが普及することによって半分は解決しようである。

5. 謝辞

日頃御指導いただき、渋谷主管研究員、CLグループの皆様には感謝いたします。

6. 参考文献

- 1) IBMマニール: STAIRS/VS GI GH12-5114.
- 2) " : STAIRS/VS-TLS GI manual GH12-5123
- 3) 粕永・小面: "カナ文字文の機械処理のための辞書について" 信託会 AL研 10A
- 4) 藤田・鶴見・首藤・吉田: "日本語文の標準形について" "
- 5) 野崎: "日本語における言語理論" 昭51年電気学会連合大会予稿
- 6) 木村: "日本語処理ソフトウェア" "
- 7) 日本情報処理開発協会: "日本語情報処理システムの研究開発" 昭51.3
- 8) 松下, 山崎, 佐藤: "漢字かを混り文変換システム" 情報処理 Vol.15 No.1.
- 9) 相川, 橋本: "自然語処理における機械辞書採索の一考察" 第17回情報処理大会予稿
- 10) 坂本: "Anaphoric approach による自動抄録への適用" "