索引作成のためのプログラムライブラリ

索引作成のためのプログラムライブラリ

中野　洋

0. まえまがき

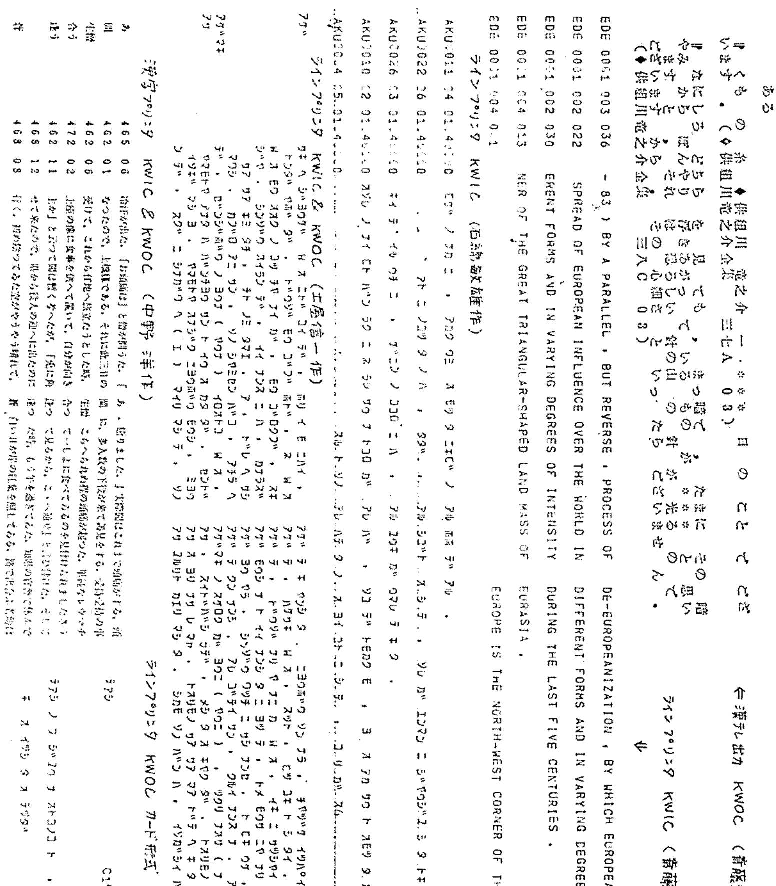
国立国語研究所に電子計算機が導入されて10年になる。この間に，言語データを電子計算機に処理させるための様々な研究がおこなわれた。これらの研究は大きく三つに分けられる。一つは，言語の自動処理の研究，一つは，分析資料作成に関する研究，一つは，言語データの分析・研究である。石綿らの構文解析の自動化の研究は自動処理の研究であり，語彙調査は言語データの分析・研究に属すると言えよう。自動処理の研究は，現在，情報検索や質問応答システムに向いつつある。これらを実現させるためには，文や文章の意味の処理が可能にならなければならない。この要求に答えるには，言語自体の分析・研究をもっと深める必要がある。そのための道具立ての一つとして，分析資料作成に関する研究に用語索引作成プログラムの開発がある。

用語索引作成プログラムは，ルーンの研究以来，最も長い歴史を持つが，国語研究所でも，斎藤秀紀（1968）「電子計算機と漢テレによる用語総索引の作成」（国研報告31「電子計算機による国語研究」）以来各種の研究・開発が重ねられて来た。また最近の言語情報処理界の成果として，植村俊亮(1975)「電子計算機による自動索引の研究（上，下)」（電子技術総合研究所報告第743，747号）があるし，出版業界においても，金田一春彦・清水功・近藤政美（1973）「平家物語総索引」（学習研究社）の電子計算機による索引作成があった。

これらの用語索引作成プログラムは，およそ次のように分けることができる。

(1) 入力

図1　文脈つき用語索引出力形式



(文字種)　カナ・英文字
　　漢字かなまじり
(媒　体)　カード
　　紙テープ
(情　報)　単位切り情報のみ
　　よみがなもつく
　　語種・品詞・活用情報もつく

(2)　出力

(文字種)　カナ・英文字
　　漢字かなまじり
(媒体)　ラインプリンタ
　　高速漢字ラインプリンタ
　　漢字テレタイプ印字機
(形　式)　KWIC
　　KWOC
　　KWIC & KWOC

入力・出力のそれぞれの組み合わせによって各種の用語索引作成プログラムが存在する（図1. KWIC 例参照)。例えば，斎藤のプログラムは，(入力文字種) 漢字かなまじり (入力媒体) 紙テープ (入力情報) 単位切り情報のみ (出力文字種) 漢字かなまじり (出力媒体) 漢字テレタイプ印字機 (形式) KWOC である。

これらのどれを選ぶかは，分析目的による。しかし，語彙調査等の進行により大量の言語データが蓄積されつつある現在，どのような形態でも出力できる機能を持ったプログラムは用意されていてよいし，これは我が国語研究所の責務でもある。

1.　目的

用語索引作成プログラムは方法論的には，先行の研究でほとんど解決されている。そこで，今回，言語計量研究部第一研究室「漱石・鷗外の用語の研究」における索引作成プログラムでは，これまでに開発された様々の機能をすべて持ったプログラムシステムを作成すること，大量調査にともなうエラーデータの処理機能を持たせることを大きな目的とした。その結果，以下に述べるようなシステムになった。

(入力文字種)　漢字かなまじり文（カナ・英文字でも可）
(入力媒体)　紙テープ（カードでも可）
(入力情報)　単位・よみがな・語種・品詞・活用情報（1つでも可）
(出力文字種)　漢字かなまじり文　または　カナ（英文字でも可）
(出力媒体)　高速漢字ラインプリンタ　または　漢テレ　または　ラインプリンタ
(形式)　KWIC & KWOC (KWOC でも可)
機能　索引作成，エラーデータの自動チェック，修正，ワードカウント
一語検索，品詞検索，

筆者は，他に，漢字かなまじり文（単位切りされていないもの）を入力として，上記の機能を持った自動処理システム（一貫処理システムと称している）を開発中である。これら二本のプログラムシステムによって，分析資料作成に関するプログラム開発は一応の完成を見たと考える。以後は，このプログラムによって作られた資料を用いて，言語の分析に向おうと考えている。

2. プログラムの説明

用語検索システムは，次の三つのサブ・システムに分れる。

1. 入力データのチェック
2. 文脈つき用語索引の作成
   2—1　データの作成
   2—2　出力

2—3　データのチェックおよび修正

2—4　異なる言語単位データ（C単位・L単位）の作成

2—5　一語検索および一品詞検索

3.　ワード・カウント

3—1　全体度数順語彙表作成

3—2　全体50音順語彙表作成

3—3　語種別集計表作成

3—4　品詞別集計表作成

2—0　文脈つき用語索引の作業手順

詳しくは，本誌1ページ～17ページの鷲岡昭夫「言語研究のための索引作成システム」に譲る。ここでは，人手作業と機械作業の組み合わせについて概略を説明する。

(1)　原文の単位切り作業　このシステムでは三種類の単位が使える。もちろん一種類でもよい。原文のコピーに直接区切り記号を書き込む。

(2)　清書・かなつけ　単位情報やよみがな情報をつけて清書する。

(3)　各種情報つけ　語種・品詞・活用・連語情報をつける。

(4)　漢テレパンチ

(5)　機械処理（入力・チェック）

(6)　校正・修正

(7)　機械処理（入力・データ作成・出力・チェック・修正）

(8)　校正・修正データの作成　(7)で作成された原文・KWICにより校正し，修正データを作成する

(9)　機械処理（修正・データ作成・出力）

(1)から(9)までは，「漱石・鷗外の用語の研究」に用いたルーチンである。(5)と(6)は新聞語彙調査システムにおいて，「データの一次パンチ→校正→修正パンチ→機械処理」の行程をふんでいたが，この「校正」の機械的な部分，すなわちフォーマットエラーや情報つけエラーのチェックを自動化したものである。

これらは(8)と重複する部分であって，エラーが少ない場合は省いてもよい。

各種情報をつけない場合は(3)を省いてよい。

カナ入力，ローマ字・アルファベット入力で情報をつけない場合は，(2，3)を省き，(4)は漢テレパンチでなくてもよい。

外国語など分かち書きをしている場合は，(1，2，3)を省いてよい。

もちろん，エラーがない場合は，(7)で終ってよい。

すなわち，これらの作業手順は，そのデータや処理の程度によって随時選択できるように設計されている。

2—1　入力データのチェック

大量データの処理においては必ずエラーデータのチェック・修正が大きな問題として取り上げられる。この問題をいかに処理するかが，そのシステムの優劣を決めるポイントとなる。

データチェックの内容は，フォーマットエラーや漢テレパンチの際に起る桁ずれエラーのような機械的なエラーから，単位切りエラーなどの単語の認定にかかわる高度なエラーまで色々ある。これが完全に機械化出来るということは，言語処理の完全な自動化に通じる。このシステムでは現在そこまではいっていないが，各種情報を用いた可能な限りのチェックを起っている。

「作業手順」に明らかなように，このチェックは必ず人間の目を通ることになっている。従って，「エラーデータであるもの」のみならず，「エラーデータらしきもの，非常に珍しいデータ」にもチェック情報をつけ，人間による校正のたしにした。

本システムのデータチェックの能力は以下の通りである。

活用形変換や50音順にするための配列情報つけのためのデータチェックは完全である。これは以後にそのデータを用いた処理があるためである。漢テレパンチに起る桁ずれエラーはそれが機械的な性質を持つため完全に発見できる。

表1はフォーマットチェック，品詞連続チェック，語形チェックによって発見されたエラーと発見されなかったエラーの状況である。

表に示すとおり，エラーの箇所311のうち，その34%，105箇所は機械的に発見することができた。このうち，フォーマットエラーは機械的な規則にはずれたものであるため，発見が容易であるが，その他のエラーの発見はむつかしい。しかし，全体の3分の1強が発見されたのであるから，このルーチンはシステムにとって有効であることがわかる。

表1　「寒山拾得」における校正状況

校正された箇所	チェック情報		計
	あり	なし	
フォーマット	38	12	50
単位情報	11	34	45
見出し語	22	31	53
よみ	7	25	32
語種情報	2	12	14
品詞情報	15	55	70
活用情報	10	37	47
計	105	206	311

(データ数　4,066)

このサブシステムは，次の四つのプログラムからなる。(　)内はプログラムＩＤである。

(1)　データの読み込み，フォーマットチェック　(NCHECK 1)

(2)　情報チェック1　(NCHECK 2)

(3)　情報チェック2　(NCHECK 3)

(4)　紙テープ打ち出し　(NCHECK 4)

以下に，各プログラムの説明をする。

2.1.1　データの読み込み，フォーマットチェック

〔処理の概要〕　読み込み……入力データ（紙テープ）を読み込む。復改記号(C/R)で区切られた部分を1レコードとして，出力レコードのフォーマット通りの固定長になおす。

フォーマットチェック……固定長になおす際に，入力データのフォーマット

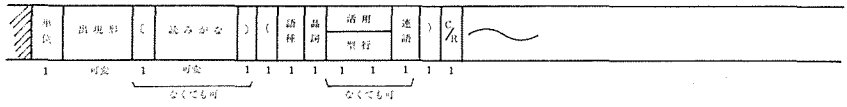
チェックと桁ずれチェック（J字チェック）をする。内容は次のとおり。

1. 先頭は，CLS（@≒以外であってはならない。
2. 〔 があれば，それ以降 C/R 以前に〕がなければならない。
3. （ があれば，それ以降 C/R 以前に）がなければならない。
4. J字チェック　漢テレ一字は2バイトで構成される。この時，前の1バイト目に $(41)_8$ のコード（J）が来てはならない。入力データの1レコードには必ず C/R がある。この C/R は＊Jで構成されており，それ以前で桁ずれがおこれば，J字チェックで検出される。

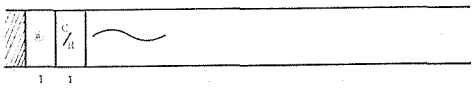
〔入力〕　紙テープ（原文）　13ページ清書例参照

フォーマット　数字は漢テレ字数を示す。

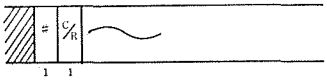
通常データ



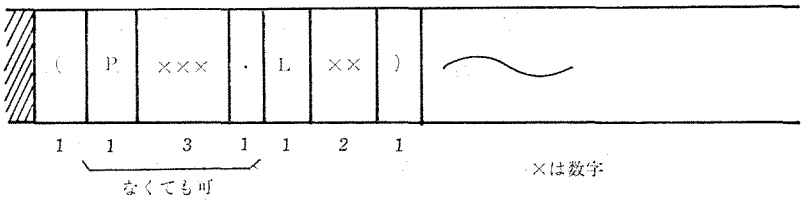
題データ　表題の語の始まりを示す



段落データ　段落の始まりを示す



頁，行データ　頁，行の始まりとその数値を示す



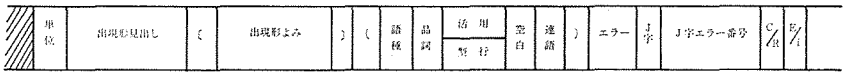
それぞれのレコードは原文順に並んでいる。ギャップは頁がかわった所に

ある。

情報コードは12ページ表１を参照のこと。

単位はＣＬＳ，それぞれ単位の始まりを示す。

〔出力〕 磁気テープ(119バイト／レコード，20レコード／ブロック，固定長)



2 40 2 40 2 2 2 2 2 2 2 2 2 4 2 8 2 1

内容はすべて3010コード。数字はバイト数。

2.1.2 情報チェック１

〔処理の概要〕 ここでの処理は三つに分れる。品詞連続に関するチェックとひらがな書きの語についての品詞情報チェック，および，連語情報の作成である。

1. 品詞連続のチェック

(1) サ変動詞語幹の後にはサ変動詞が来ること。

勉強／する　リード／する

サ変動詞の前に必ず，サ変動詞語幹がくるとはかぎらない。

びっくり／する　一／周／する

(2) 形容動詞語幹の後には助動詞「だ」，または名詞性接辞「さ・み・げ」がくる

静か／な　あわれ／み

Ｃ単位は二つのＬ単位で構成される。前のＬ単位は自立語部分に相当し，後のＬ単位は附属語部分に相当する。前のＬ単位内にあらわれた助動詞は形容動詞の語尾に相当するものである。この助動詞の前に必ず形容動詞語幹がくるとはかぎらない。(田中章夫執筆 (1972)「形容動詞の諸形態」国研報告42「電子計算機による新聞の語彙調査Ⅲ」参照)

幾何／学／的／な　減り／ぎみ／で　苦し／げ／に　起こし／がち／

で　スケッチ／風／な　ステレオ／向き／な　うれし／そう／な

無／軌道／な　有／意義／な　急／ピッチ／な

(3)　形容詞語幹・形容詞派生形の後には助詞「の」名詞性接辞「さ，み，げ」がくる。

なつかし／の／メロディ　さみし／げ　うれし／さ

2.　ひらがな書きの語についての品詞情報チェック

新聞の語彙調査で，上位100語をとれば，それだけで全体の44.4%，200語をとれば49.0%（記号を含む）をしめる。よく出現する語は辞書に入れてチェックした方がよい。上位100語，200語といっても，新聞だけによく出現する語や，ある分野だけによく出現する語が入る。それを除くために，ここでは，ひらがな書きの語だけを上位200語とり，辞書に収めた。ひらがな書きの語にしたのは，そこに基礎語彙が多くはいるだろうことと，入力データのすべてについてこの辞書をひくという非能率を避けるためである。

辞書に収めた語のうち，同形異語のある場合は当然いくつかの情報もつけておかなければならない。この時，どちらの話がより多く出現するかを考えなければならない。たとえば，「し」については，「する」「死」「四」などが考えられる。しかし，語形「し」では，そのほとんどが「する」の活用形であり，「死」や「四」を「し」とあらわすことは少ないだろう。この時，辞書を (SE, T1, T7) としておくと，(SE) のつもりで，(T1 や T7) と書きまちがえた時のチェックにはならない。辞書を (SE) だけにしておけば，上のような間違いはチェックできるし，本当に (T1, T7) の時に，チェック情報がついても，何回も起ることはないのだからかまわないだろう。逆に，校正作業者がチェック情報にたより切るという弊害をなくす働きもするだろう。

今回，おこなったチェック用辞書に収められている語数は300語である。辞書の順序は，新聞語彙調査における度数順にした。ただし，今回の調査の付加情報規則 (LDP 10 参照) に定めた，動詞性接辞，形容詞性接辞，助動

詞，助詞などもこの辞書に入れ，後につけた。

3. 連語情報の作成

- 連語情報がC単位内のどのS単位についていても同じC単位内の全てのS単位に同じ情報をつける。
- 品詞情報が，サ変語幹である語とその次の語に連語情報「サ」をつける。
- 形容動詞語幹あるいは名詞性接辞でそれが接辞辞書内の語（的，風など）のつぎの語が助動詞であれば，それぞれに連語情報「ケ」をつける。
- 副詞のつぎが副詞語尾であれば，それぞれに連語情報「フ」をつける。

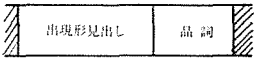
〔入力〕　磁気テープ

フォーマット　NCHECK1の出力磁気テープと同じ

紙テープ1（チェック用テーブル　ひらがな書きの語の品詞辞書）

紙テープ2（形容動詞の語幹に用いる接辞の辞書）

フォーマット　紙テープ1



紙テープ2

数字は洩テレ字数

〔出力〕　磁気テープ

フォーマット　NCHECK 1 の出力磁気テープと同じ

2.1.3 情報チェック2

〔処理の概要〕　このプログラムでは，ページや行情報が上昇順になっているかどうか，語種と品詞情報の組み合わせが正しいかどうかを調べる。

たとえば，漢語や外来語は名詞が多く，動詞や形容詞，連体詞などはない（「感ずる・デモる」などは混種語）。漢語の副詞はあるが，外来語の副詞はない。英語の接続詞・感動詞などは例数が少なく，珍しいデータといえる。このようなことは「電子計算機による新聞の語彙調査Ⅱ」（国研報告38）の「語彙量の分析――語種・品詞別の異なり語数分布」にあきらかである。

今回のチェックでは，次の組み合わせ以外のものにチェック情報をつけた。

S1, S3, S4, S5, S6, S7, S9, SA, SB, SC, SD, SE, S+, S−, SM,

T1, T3, T4, T6, T7, TC, U1, U3, U4, U6, U7, V1, V3, V4, VC VE, VM, W8, WP, WR, X7, XX, YY (本誌12ページ付加情報コード表参照のこと)

〔入力〕 磁気テープ

フォーマット 情報チェック1と同じ。

紙テープ

フォーマット 語種 (漢テレ1字) ／品詞 (漢テレ1字)

〔出力〕 磁気テープ (119バイト／レコード, 20レコード／ブロック, 固定長)

フォーマット 入力磁気テープと同じ

2.1.4 紙テープ打ち出し

〔処理の概要〕 入力磁気テープを読み, フォーマット通りに紙テープを出力する。

〔入力〕 磁気テープ (119バイト／レコード, 20レコード／ブロック, 固定長)

フォーマットは NCHECK 1 の出力磁気テープと同じ。

〔出力〕 紙テープ (1ページ分／1ブロック, 可変長)

フォーマットはNCHECK 1 の入力にほぼ同じ。 (C/Rの前にエラーまたは警報記号が出力される)

2.2 文脈つき用語索引の作成

ここでは, 紙テープまたはカードにパンチされた入力データをよみこみ, KWICレコードを作る「データ作成」, 漢字プリンタ・ラインプリンタまたは, 漢字テレタイプ印字機に出力する「出力」, 磁気テープ内に納められた「データのチェックおよび修正」, 指定された語または品詞を取り出す「一語検索および一品詞検索」のそれぞれのプログラムについてのべる。

「データ作成」では, もともと可変長である言語データを固定長レコードに直し, 以後の処理を容易にすること, 検索しやすくするために活用語を代表形

に変換すること，五十音順に並べるための配列情報をつくること，および各語の用例を作ることが主な処理になる。それぞれの処理をデータにあわせて自由に選択できるようにするため，レコードのフォーマットは統一される。また，漢字ラインプリンタにでもラインプリンタにでも出力できるようにするため，漢字かなまじり文とカナ文の二種の用例を作る。「漱石・鷗外の用語の研究」では，三種類の言語単位を採用した。それぞれの語が見出しとなるレコードも作成する。作られたレコードはワードカウントの入力データとなる。

「データのチェックおよび修正」では，機械による自動チェックをおこなうこと， KWIC 等を利用した人間によるチェックの結果，発見されたエラーレコードを単語単位で修正することが主な処理である。

それぞれのプログラムは次の通りである。

データ作成

(1) 紙テープ（原文）読み込み，フォーマットチェック (NINPUTSKN)

(2) フォーマット変換，出典情報・文種・題情報・段落番号の作成 (NDATASAKU)

(3) 活用形変換 (NKATSUYO)

(4) 配列情報つけ (NGOJYUON)

(5) コピー (DUP)

(6) かな用例つけ (NYOUREI 0)

(7) 漢字かなまじり用例つけ (NYOUREI 1)

(8) 併合 (NMERGESKN)

(9) 50音順ソート (SORT)

(10) カード（原文）読み込み (NINPUTCR)

(11) 紙テープ（かな原文）読み込み (NINPUTPT)

(12) L単位・C単位データの作成 (NLCTANI)

出力

(13) ラインプリンタ（かな）出力 (NOUTPUT 1)

(14) 漢字プリンタ出力用編集 (NOUTPUT 2)

(15) 漢テレ出力 (NOUTPUT 3)

(16) 原文漢字プリンタ出力用編集 (NTEXTSKN)

(17) 漢字プリンタ出力 (CVT-MT, MT-HKP)

データのチェックおよび修正

(18) データチェック (NCHECKMT)

(19) 紙テープ（エラーデータ）打ち出し (NCHECKOUT)

(20) 修正パラメータ作成 (NPARAM 1)

(21) データの修正 (NSYUSEI MT)

(22) フォーマット変更 (NKATAGAE)

一語検索および一品詞検索

(23) 一語検索 (NGOR)

(24) 一品検索 (NHINSHIR)

次にそれぞれのプログラムの説明をする。

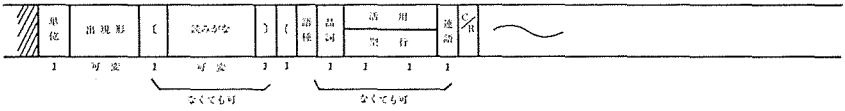
2.2.1 紙テープ（原文）読み込み，フォーマットチェック

〔処理の概要〕 このプログラムでは，紙テープ（原文）データを読み込み，出力フォートマットどおりに編集して磁気テープ出力する。なおこの時，入力紙テープがフォーマット通りかどうかのチェックもあわせ行う。(NCHCK 1 とほぼ同じ内容)

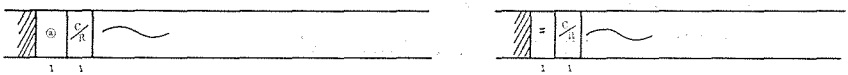
〔入力〕 紙テープ（可変長・最大五千漢テレ字・1ページ1ブロック)

フォーマット

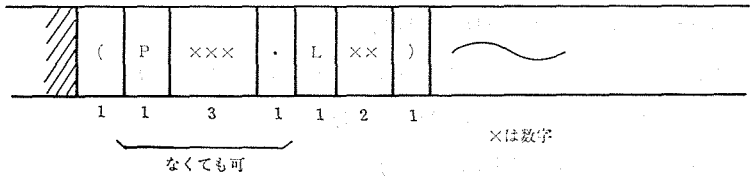
通常データ



題データ　表題の語の始まりを示す　　段落データ　段落の始まりを示す



頁，行データ　　頁，行の始まりとその数値を示す



頁・行データは頁や行がかわった所に入れる。

題データは文章の章や節の見出しの初めに入れる。段落データで解消される。

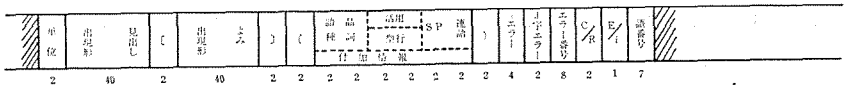
段落データは段落の初めに入れる。

単位情報・付加情報のつけ方およびコードは，インプットデータのチェックのプログラム・NCHECK 1の項参照。

〔出力〕　磁気テープ(固定長・126バイト／レコード，20レコード／ブロック)

フォーマット

通常データ



頁・行データは）・C/R 以外は先頭から順につまる。

題・段落データは，・C/R 以外は先頭から順につまる。

2.2.2　フォーマット変換，出典情報・文種・題情報の段落番号の作成

〔処理の概要〕　このプログラムでは，NINPUTSKN（または，NSYUSEI MT）の出力磁気テープを入力データとしてフォーマット変換，出典・情報・

文種・題情報・段落番号の作成，および分類語彙表の番号の付加をおこなう。処理の内容は次の通り。(～はスペースを表わす)

(1) 単位情報　C→ＣＬＳ，Ｌ→～ＬＳ，Ｓ→～～Ｓ（共にEBCDIKに）

(2) 〔，〕，(，)，Ｊ字エラー，Ｊ字エラーの NO削除。

(3) 語種・品詞・活用の型を，漢テレコードからEBCDIKコードへ。

(4) 活用の行を，下記のコード (EBCDIK) へ。

ワ→0，あ→1，か→2，が→3，さ→4，ざ→5

た→6，だ→7，な→8，は→9，ば→A，ぱ→B

ま→C，や→D，ら→E，わ→F，V→G，Z→G

(5) 連語情報を下記のコード (EBCDIK) へ。

サ→3，ケ→4，コ→8，セ→A，カ→B，フ→C，レ→D

ド→E，ヨ→L

(6) 出典情報の作成　全てのレコードに，Key In された作品名をつける。次に新しく変わるまでの全てのレコードに，頁・行データの数字部分を入れる。

(7) 文種情報の作成　出現形見出しが，「であるデータから，」であるデータの直前までの全てのレコードに文種情報（K）をおくる。

それ以外のレコードには，文種情報（G）をおくる。

(8) 題情報の作成　題データ（@）からはじまって，次の段落データまでのすべてのレコードに題情報（@）をおくる。それ以外のレコードには，題情報（G）をおくる。題データは削除する。

(9) 段落番号の作成　最初の段落データを１とし，段落データがくるまでの全てのレコードにおくる。段落データは削除する。

(10) 磁気ディスクの総合辞書によって分類番号をつける。

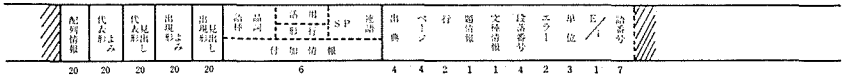
〔入力〕　磁気テープ(固定長，126バイト／レコード，20レコード／ブロック)フォーマットは，NINPUTSKN の出力磁気テープに同じ。

磁気ディスク（総合辞書，固定長，105バイト／レコード，ISAM

ファイル)

〔出力〕　磁気テープ（固定長，135バイト／レコード，20レコード／ブロック）

フォーマット



配列情報，代表形よみ，代表形見出し欄にはスペースを入れる。

入力データの出現形よみ・出現形見出しが20バイト以上あれば，18，19バイト目に＊（漢テレ字）を入れる。

ラインプリンタ（エラーリスト）

エラーレコード

2.2.3　活用変換

〔処理の概要〕　動詞・形容動詞活用レコードについて，活用情報を用いて終止形に変換する。このプログラムは鷲岡昭夫氏の作成によるものであり，処理内容については，鷲岡昭夫「文語形・口語形活用語の代表形変換について」（電子計算機による国語研究Ⅴ）に詳しい。

〔入力〕　磁気テープ(固定長，135バイト／レコード，0レコード／ブロック)

フォーマットは，NDATASAKUの出力磁気テープと同じ。

紙テープ（活用変換テーブル)

フォーマットは，鷲岡論文参照。

〔出力〕　磁気テープ(固定長，135バイト／レコード，20レコード／ブロック)

フォーマットは，入力磁気テープと同じ。

活用レコードは，その終止形が代表形よみ・代表形見出しに入る。それ以外のレコード（活用エラーレコードを含む）は，出現形よみ・出現形見出しが代表形よみ・代表形見出しに入る。

ラインプリンタ（エラーリスト）

エラーレコード

2.2.4 配列情報つけ

〔処理の概要〕 このプログラムでは，代表形よみについて，50音順にレコードを並べるための配列情報を作成する。このプログラムは田中章夫氏の作成による。処理の内容については，田中章夫「電子計算機によるワードリスト作成上の一問題」（電子計算機による国語研究）に詳しい。

〔入力〕 磁気テープ（固定長135バイト／レコード，20レコード／ブロック）
フォーマットは，NDATASAKUの出力磁気テープと同じ。

紙テープ（かな TABLE)
フォーマットは，田中論文参照。

〔出力〕 磁器テープ（固定長，135 バイト／レコード，20 レコード／ブロック フォーマットは，入力磁気テープと同じ。

ラリンプリンタ（エラーリスト）

エラーレコード

2.2.5 コピー

〔処理の概要〕 このプログラムでは，入力磁気テープと全く同じ内容の磁気テープを作り，出力する。サービスルーチンを用いる。

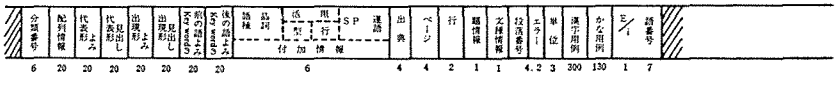
〔入力・出力〕 磁気テープ（固定長，135 バイト／レコード，20レコード／ブロック)
フォーマットは NDATASAKU の出力磁気テープと同じ。

2.2.6 かな用例つけ

〔処理の概要〕 このプログラムでは，かなの用例（130字， KWIC）を作成し出力する。

〔入力〕 磁気テープ二ファイル（固定長，135 バイト／レコード，20 レコード／ブロック)
フォートマットは， NDATASAKU の出力磁気テープと同じ。

〔出力〕　磁気テープ（固定長，611バイト／レコード，20レコード／ブロック）



漢字用例には　が入っている。

2.2.7　漢字かなまじり文用例つけ

〔処理の概要〕　このプログラムでは，漢字かなまじり文用例（300字，KWIC）を作成し，出力する。

〔入力〕　磁気テープ二ファイル（固定長，135バイト／レコード，20レコード／ブロック）

フォーマットは，NDATASAKUの出力磁気テープと同じ。

〔出力〕　磁気テープ（固定長，611バイト／レコード，20レコード／ブロック）

フォーマットはNYOUREI0と同じ。

かな用例には　が入っている。

2.2.8　併合

〔処理の概要〕　このプログラムは，NYOUREI0とNYOREI1の出力磁気テープを入力とし，かな用例と漢字かなまじり文用例をあわせもつレコードを作り，磁気テープ出力する。

〔入力・出力〕　磁気テープ（固定長，611バイト／レコード，5レコード／ブロック）

フォーマットは，NYOUREI0と同じ。

入力・出力とも同じ。

2.2.9　50音順ソート

〔処理の概要〕　このプログラムでは，レコードを50音にならべかえる。

ソート・キーは，次の通り。

1. 配列情報　　20バイト
2. 代表形よみ　　20
3. 代表形見出し　　20
4. 付加情報　　4（語種，品詞，活用）
5. 出現形よみ　　20
6. 出現形見出し　　20
7. Key Word の後の語の出現形よみ　　20
8. Key Word の前の語の出現形よみ　　20
9. ページ　　4
10. 行　　2

〔入力・出力〕　磁気テープ（固定長，611バイト／レコード，5レコード／ブロック）

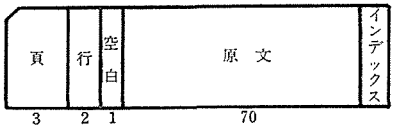
フォーマットは，NYOUREI 0 と同じ。

入力・出力とも同じ。

2.2.10 カード（原文）読み込み

〔処理の概要〕　このプログラムでは，IBM カードにパンチされた原文（カナでもアルファベットでも可，分かち書きされていること）を読み込み，一語単位で磁気テープに固定長出力される。この際，カードにパンチされた頁・行情報を出力の所定の欄に入れる。また，語は見出し語・代表形欄にそのまま送る。

〔入力〕　カード（固定長，1レコード／ブロック）



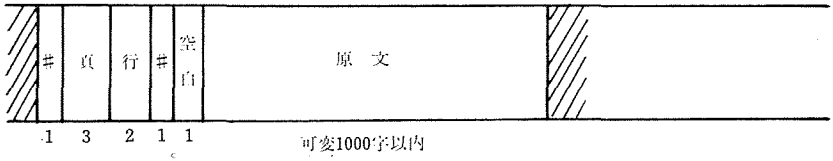
〔出力〕　磁気テープ（固定長，135バイト／レコード，20レコード／ブロック）

フォーマットは，NDATASAKU の出力磁気テープと同じ。

2.2.11 紙テープ（かな原文）読み込み

〔処理の概要〕 このプログラムでは，フレキソタイプライタで打たれた，分かち書きカナ原文（アルファベットでも可，可変長）を読み込み，一語単位で磁気テープに固定長出力される。この際，紙テープ先頭にパンチされた頁・行（または識別コード）情報を出力の所定の欄に入れる。語は見出し語・代表形欄にそのまま送る。

〔入力〕 紙テープ（可変長，最大1010字以内）



〔出力〕 磁気テープ（固定長，135バイト／レコード，20レコード／ブロック）

フォーマットは，NDATASAKU の出力磁気テープと同じ。

2.2.12 L単位・C単位データの作成

〔処理の概要〕 このプログラムでは，NGOJYUON の出力磁気テープを入力として，コンソールからの Key-in によりL単位，またはC単位のレコードを作成し，磁気テープに出力する。

このプログラムで作成されたデータを入力データとして，プログラム 205～215により，それぞれの文脈つき用語索引を作成することができる。また，語彙調査用プログラムによって，それぞれの語彙表も作成することができる。

〔入力〕 磁気テープ（固定長，135バイト／レコード，20レコード／ブロック）

フォーマットは，NDATASAKU の出力磁気テープと同じ。

〔出力〕 磁気テープ（固定長，135バイト／レコード，20レコード／ブロック）

フォーマットは，入力磁気テープと同じ。

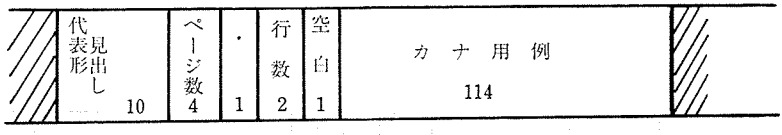
2.2.13 ラインプリンタ（かな）出力

〔処理の概要〕 このプログラムは，NSORTの出力磁気テープ（索引ファイル）を入力として，かな文の文脈つき用語索引をラインプリンタに出力する。

〔入力〕 磁気テープ（固定長，611バイト／レコード，5レコード／ブロック，フォーマットは，NY UREI 0 と同じ。

〔出力〕 ラインプリンタ（132字／行）

フォーマット



印字例

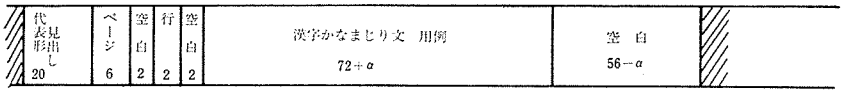
カケル　0465・03　マ　ニ　フキ　カケタ　。　リヨ　ハ

2.2.14 漢字プリンタ用出力編集

〔処理の概要〕 このプログラムでは，NSORTの出力磁気テープ（索引ファイル）を入力として，漢字かなまじり文の文脈つき用語索引を漢字ラインプリンタに出力するための磁気テープを作る。漢字プリンタ1ページに用例42行，ほかに作品名，ページ2行を出力する。ただしA4版，1行53字詰。

〔入力〕 磁気テープ(固定長，611バイト／レコード，5レコード／ブロック)フォーマットは，NYOUREI 0 と同じ。

〔出力〕 磁気テープ(固定長，160バイト／レコード，1レコード／ブロック)
漢プリ一行分を1レコードとする



αは盤外字数×2。盤外字記号＋漢字2字を一字で印字するため。

印字例

間　462　1　である。それに此三日の　間　に，多人数の下役

2.2.15　漢テレ出力

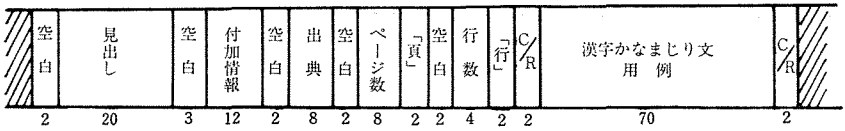
〔処理の概要〕　このプログラムは，NSORTの出力磁気テープ（索引ファイル）を入力として，漢字かなまじり文の文脈つき用語索引を紙テープで出力する。この紙テープは漢テレ印字機にかけられる。

〔入力〕　磁気テープ(固定長，611バイト／レコード，5レコード／ブロック)

フォーマットは，NYOUREI 0 と同じ。

〔出力〕　紙テープ（固定長，144バイト／レコード，1レコード／ブロック）

フォーマット



印字例

間　　S 1　　寒山拾得　　462頁　　1行

…である。それに此三日の　間　に，多人数の下役が来て謁見をする…

2.2.16　原文出力

〔処理の概要〕　このプログラムでは，配列情報つけ（NGOJYUON）済みファイルをもとにして，原文イメージの出力（漢字プリンタ用編集）磁気テープを作る。漢字プリンタ1ページに原文行，ほかに作品名，ページ2行を出力する。ただし，A4版，1行53字詰。

〔入力〕　磁気テープ(固定長，611バイト／レコード，5レコード／ブロック)

漢字プリンター行分を1レコードとする。

フォーマット　先頭から53文字分（盤外字詰記号＋漢字2字を一字に数える）データをつめ，残りは漢テレ　を送る。

2.2.17 漢字プリンタ出力

〔処理の概要〕　このプログラムは，漢字プリンタ出力用に編集された磁気テープを入力とし，まず，CVT-MT80 で国研漢テレコードを日電漢プリコードにコンバートし，次に MT-HKP で漢字プリンタに印字出力する。

これらのプログラムは，漢字プリンタ出力用に用意された汎用プログラムである。

2.2.18 データチェック

〔処理の概要〕　このプログラムでは，NINPUTSKN の出力磁気テープを入力として，データのチェックをおこなう。チェックの内容は下記の通り。

エラーデータはラインプリンタと磁気テープに出力される。

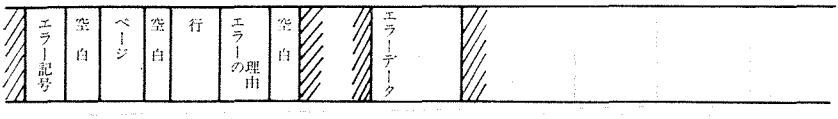
1. J字チェック　NINPUTSKN のチェックの結果を利用する。
2. ページ，行のチェック　だぶり，とばしがあるかどうか，桁オーバも調べる。
3. 見出し，よみがなの桁チェック　漢テレ字で11字以上のものはエラーデータとして出力する。
4. 活用の型・行情報のチェック　コード表以外の記号がくれば，エラーデータとする。
5. 連語情報のチェック　コード表以外の記号がくれば，エラーデータとする。

〔入力〕　磁気テープ（126バイト／レコード，20レコド／ブロック，固定長）

フォーマットは，NINPUTSKN の出力と同じ。

〔出力〕　ラインプリンタ（132バイト，固定長）

1レコード　2行



エラーデータは，漢テレコードは，3010 コードの印字形式で EBCDIK コードに変換して印字する。

磁気テープ（126 バイト／レコード，20 レコード／ブロック，固定長）

フォーマットは，入力と同じ。エラーデータだけを，そのまま，出力する。

2.2.19 紙テープ打ち出し

〔処理の概要〕 データチェックの出力磁気テープを入力とし，NINPUTSKN の入力紙テープのフォーマット通りに出力する。

この紙テープは，漢テレ印字機にかけられ，エラーデータの修正に用いられる。

〔入力〕 磁気テープ（126 バイト／レコード，20 レコード／ブロック／固定長）

フォーマットは，NINPUTSKN の出力と同じ。

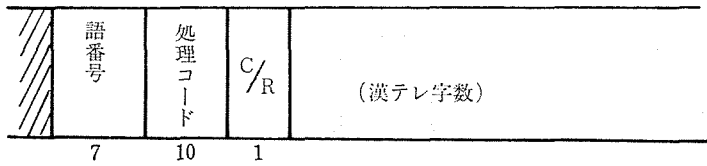
〔出力〕紙テープ（可変長，ページ情報データまでを 1 ブロックにする。）

フォーマットは，NINPUTSKN の入力と同じ。

2.2.20 修正パラメータ作成

〔処理の概要〕 このプログラムでは，修正パラメータを入力とし，このパラメータの順序が正しいかどうかをチェックし，正しければ，磁気テープに出力し，誤っていれば出力しない。ラインプリンタに入力データの語番号を出力し，それぞれの修正の方法とチェックの結果を出力する。

〔入力〕 紙テープ（修正パラメータ） 2 ブロックで 1 レコードとする。フォーマットは，前のブロックのデータは下記の通り，後のブロックのデータは，NINPUTSKN の入力と同じ。



語番号は漢テレ7字，処理コードは，　の時，削除。コードがない時は，さしかえか挿入。修正されるデータに同じ語番号があれば，さしかえ，なければ，挿入。

〔出力〕　磁気テープ（126バイト／レコード，20レコード／ブロック，固定長）

フォーマットは，NINPUTSKN に同じ。

削除レコードは，語番号以外はスペース。

ラインプリンタ（132バイト）二行で一レコード

	PARAM	空白	パラメータのデータ	
	サシカエ ソウニュウ	空白	修正済データ	

チェック欄には，エラーデータのみ，ERROR WORD NUMBER と出力する。

修正方法の欄には，削除の時のみ，CUT と出力する。

2.2.21　データの修正

〔処理の概要〕　このプログラムでは，NINPUTSKN の出力磁気テープを入力として，そのエラーデータを NPARAM 1 の出力磁気テープによって修正する。

〔入力〕　磁気テープ　2ファイル（126バイト／レコード，20レコード／ブロック）

フォーマットは，NINPUTSKN の出力と同じ。

〔出力〕　磁気テープ（126バイト／レコード，20レコード／ブロック）

フォーマットは入力と同じ。

入力の二ファイルの語番号が同じなら，修正パラメータにさしかえする。この時，修正パラメータの内容が語番号以外スペースなら削除する。語番号が異なれば，修正パラメータを挿入する。

2.2.22 フォーマット変更

〔処理の概要〕 このプログラムでは，配列情報つけ済のデータを紙テープ読み込み済のデータにフォーマット変更をおこなう。これは，修正されたデータをマスターテープにするために必要である。

〔入力〕 磁気テープ（固定長，135バイト／レコード，20レコード／ブロック）

フォーマットは，NDATASAKU の出力と同じ。

〔出力〕 磁気テープ（固定長，126バイト／レコード，20レコード／ブロック）

フォーマットは，NINPUTSKN の出力と同じ。

2.2.23 一語検索

〔処理の概要〕 このプログラムでは，NSORT の出力磁気テープ（索引ファイル）を入力として，漢テレで打たれた紙テープにより指定された語を抜き出し，磁気テープに出力する。

このプログラムの出力磁気テープは，NOUTPUT 1，2，3によりそれぞれ印字出力される。

〔入力〕 磁気テープ(固定長，611バイト／レコード，5レコード／ブロック)

フォーマットは，NYOREI 0 と同じ。

〔出力〕 入力磁気テープと同じ。

2.2.24 一品詞検索

〔処理の概要〕 このプログラムでは，NSORT の出力磁気テープ（索引ファイル）を入力として，コンソールからの Key-in により指定された品詞を抜き出し，磁気テープに出力する。

このプログラムの出力磁気テープは，NOUTPUT 1，2，3によりそれぞれ印字出力される。

〔入力〕 磁気テープ(固定長，611バイト／レコード，5レコード／ブロック)

フォーマットは，NYOUREI 0 と同じ。

〔出力〕　入力磁気テープと同じ。

2.3　ワード・カウント

ここでは，このシステムで作られたデータの分析のために，各種の語彙表・集計表をつくる。

本来，用語検索のために作られたデータは，語彙調査に用いないほうがよい。というのは，用語検索用の言語単位と語彙調査用の言語単位は異なるのが普通だからである。この点については，本誌鷲岡論文に詳しいが，筆者の考えを簡単に述べる。

まず第一の理由は用語検索と語彙調査とは目的が違うということである。用語検索は検索しやすいことを第一の目的としなければならない。したがって，単位は短いほうがよい。この点では文字単位が最もよいことになるが，こうするとデータが多くなり検索速度が落ちる。したがって，形態素レベルの単位が用語検索用の単位として適当ではないかと考える。一方，語彙調査はそれ自体言語学的な目的を持っている。その目的にあった言語単位を用いるべきである。

第二に，用語検索用データにつけられた品詞情報はその単位が短いため，情報つけのレベルが異なる。たとえば，接辞や語幹などは造語成分であり，動詞・形容詞……などは自立語成分である。この中間的なものとして，接辞ではないが，自立しないもの（合理，具体，本格など）も存在する。接続詞などにはそれ自体で接続詞の機能を持つものもあるが，いくつかの語がつながって接続詞の機能を発揮するもの（そうして，というのは等）もある。このような情報を一まとめにしてワードカウント用に扱うのは問題が残る。

第三に，データの採集範囲が問題になる。たとえば，文学作品を例にとると，用語検索用データにはまえがきや表題，注なども入っていてもよいし，その方が便利であるが，語彙調査用データには入っていない方がよい。

第四に，同語異語判別の問題がある。用語検索用データは必ずしも同語異語判別されていなくてもよい。むしろ，表記や語形変化を調べるためにはしてい

ない方がよい。語彙調査では同語異語判別がされていないとこれが致命的な欠陥となる。

以上のような理由で用語検索用データを語彙調査に用いない方がよいが，それでもこのサブシステムを設けたのは次の理由による。

その第一の理由は，データ量の大まかな数値を得ることは意味がある。第二は，用語検索用データを語彙調査用に加工した場合に，このシステムは有用である。第三に，逆にこのシステムを語彙調査用に用いることができる。語彙調査において最も重要なのは同語異語判別であるが，これには KWIC を使うのが便利である。

このサブシステムは，次の四つの機能を持っている。全体度数順表作成，全体50音順表作成，語種別集計表作成，品詞別集計作成。

プログラムは次の通りである。

(1) 50音順ソート (SORT)

(2) マージ・カウント (NCOUNT)

(3) 度数順ソート (SORT)

(4) 比率計算 (NKEISAN)

(5) 全体度数順表作成 (NPRINT)

(6) 語種別集計表作成 (NSYUUKEI 1)

(7) 品詞別集計表作成 (NSYUUKEI 2)

(8) 抜き出し (NPICKUP)

(9) 漢字プリンタ出力用編集 (NKLPPRINT)

(10) 漢字プリンタ出力 (CVT-MT, MT-HKP)

2.3.1 50音順ソート

〔処理の概要〕 NGOJYUON の出力磁気テープを入力として，50音順ソートをする。

ソートキーは次のとおり。

1　配列情報　　20バイト

2　代表形よみ　　20バイト

3　代表形見出し　　20バイト

4　語種　品詞　活用情報　　4バイト

5　分類番号　　6バイト

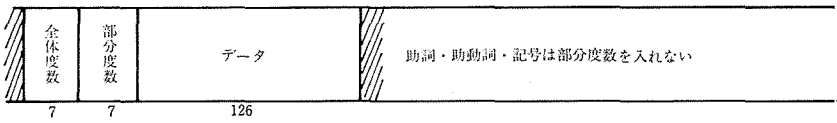
2.3.2　マージ・カウント

〔処理の概要〕　1の50音ソート済の磁気テープを入力として，ソートキー1～5をマージ・カウントのキーとして（同じものを一語と数える），のべ語数と異なり語数の計算をする。

〔入力〕　磁気テープ（135バイト／レコード，20レコード／ブロック）

フォーマットは，NDATASAKU の出力と同じ。

〔出力〕　磁気テープ（140バイト／レコード，20レコード／ブロック）



ディスプレイ

NOB　XXXXXXX　　NOBB　XXXXXXX

KOT　XXXXXXX　　KOTB　XXXXXXX　X：数字

2.3.3　度数順ソート

〔処理の概要〕　NCOUNT の出力磁気テープを入力として，その度数をソートキーにして，度数順に並べる。

2.3.4　比率計算

〔処理の概要〕　マージ・カウントの出力，全体ののべ，ことなり，部分ののべ，ことなり数をコンソールから Key-in し，各レコードの比率（出現率）と順位を計算する。計算結果は磁気テープに出力する。

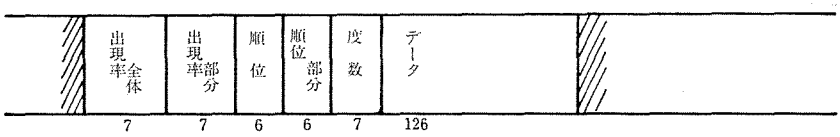
ここで，部分とは，助詞・助動詞・記号を除いたものをいう。

〔入力〕　磁気テープ（140バイト／レコード，20レコード／ブロック）

フォーマットは，NCOUNTの出力と同じ。

コンソールから，全体ののべ，ことなり，部分ののべ，ことなりを7桁の数字で入れる。

〔出力〕　磁気テープ（159バイト／レコード，20レコード／ブロック）



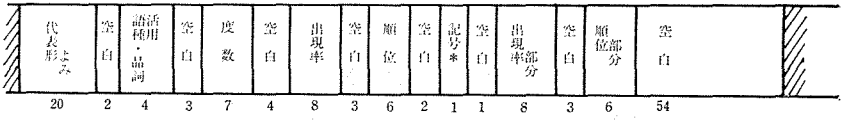
2.3.5　全体度数順表作成

〔処理の概要〕　比率計算（NKEISAN）の出力磁気テープを入力として，度数順表を作成プリントする。

〔入力〕　磁気テープ（159バイト／レコード，20レコード／ブロック）

フォーマットは，NKEISANの出力と同じ。

〔出力〕　ラインプリンタ（132バイト）



記号は（＊）部分につける。

2.3.6　語種別集計表作成

〔処理の概要〕　NPICKUPにより，語種を指定して抜き出されたデータ（磁気テープ）を入力として，語種別集計表を作成プリントする。内容はNPPINTに同じ。

2.3.7　品詞別集計表作成

〔処理の概要〕　NPICKUPにより，品詞を指定して抜き出されたデータ（磁気テープ）を入力として，品詞別集計表を作成プリントする。内容は，

NPRINT に同じ。

2.3.8 抜き出し

〔処理の概要〕 度数順配列済の磁気テープを入力とし，そこから，コンソールから Key-in された情報（語種，品詞）をもつレコードを抜き出し，磁気テープに出力する。

〔入力〕 磁気テープ（140バイト／レコード，20レコード／ブロック）

フォーマットは，NCOUNT の出力と同じ。

コンソールから，情報（語種，品詞）コードを Key-in する。

〔出力〕 磁気テープ 入力と同じ。

ディスプレイ

NOB XXXXXXX NOBB XXXXXX

KOT XXXXXXX KOTB XXXXXX X: 数字

2.3.9 漢字プリンタ出力用編集

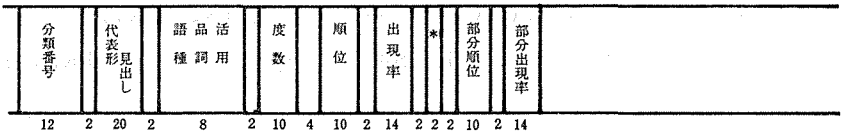
〔処理の概要〕 このプログラムでは，比率計算 (NKEISAN) の出力磁気テープを入力として，度数順表を漢字プリンタに出力するための磁気テープを作成する。出力される文字種はすべて漢テレコードにする。

〔入力〕 磁気テープ（159バイト／レコード，20レコード／ブロック）

フォーマットは，NKEISAN の出力と同じ。

紙テープ（50漢テレ字以内で表のヘッダーをいれる）

〔出力〕 磁気テープ（118バイト／レコード，1レコード／ブロック）



2.3.10 漢字プリンタ出力

〔処理の概要〕 漢字プリンタ汎用プログラム (CUT-MT, MT-HKP) を用

いて，語彙表を漢字プリンタに出力する。

後に，プログラム一覧表とフローチャートを示す。詳細は，言語計量研究部第一研究室にある仕様書類，ソースプログラムリストを参照されたい。オペレート時には，ランブックを参照されたい。プログラムは国立国語研究所の計算機 HITAC-8250 の私用実行型式ライブラリに登録してあり，ジョブカードは用意されている。また，プログラムはCOBOL で書かれており，他機種の計算機でも実行できるはずである。

3. プログラム使用例

このプログラムライブラリは，いかなる形式の入力データでも処理し，KWIC を作成し，必要ならワードカウントをおこなえるように設計されている。ここでは，その使用例を入力データの種類別に示す。

3.1 カナまたは英文字，分かち書き，カード入力

〔入力例〕 00101 ONCE WHEN I WAS SIX YEARS OLD I SAW A MAGNIFICENT...

先頭5桁頁行情報，7桁目から77桁目まで原文。一語が二枚のカードにわたってはならない。

〔使用プログラム〕 カード読み込み (NINPUTCR)→コピー (DUP)→かな用例つけ (NYOUREI 0)→50音順ソート (SORT)→ラインプリンタ出力 (NOUTPUT 1)

カード読み込み (NINPUTCR)→マージカウント (NCOUNT)→度数順ソート (SORT)→比率計算 (NKEISAN)→全体度数順表作成 (NPRINT)

〔出力例〕

```
ﾐﾀﾞｼ        ﾍﾟｰｼﾞ,ｷﾞｮｳ                    ﾖｳｼｲ                              PAGE-0014
A           008.05 ? ) MY DRAWING WAS NOT  A  PICTURE OF A HAT , IT WAS A PI
            008.05 G WAS NOT A PICTURE OF  A  HAT , IT WAS A PICTURE OF A BO
            012.16 IF HE WERE SPEAKING OF  A  MATTER OF GREAT CONSEQUENCE : '
            012.10 GAVE ANY SUGGESTION OF  A  CHILD LOST IN THE MIDDLE OF TH
            007.03  , IT WAS A PICTURE OF  A  BOA CONSTRICTOR IN THE ACT OF
            008.06  , IT WAS A PICTURE OF  A  BOA CONSTRICTOR DIGESTING AN E

ﾐﾀﾞｼ                ｼﾞﾖｳﾎｳ ﾄﾞｽｳ      ﾋﾘﾂ        ｼﾞｭﾝｲ      ﾌﾟﾌﾟﾝﾋﾘﾂ  ﾌﾟﾌﾟﾝｼﾞｭﾝｲ

,                          219     52.2673      1  %  52.2673          1
.                          199     47.4940      2  %  47.4940          2
THE                        152     36.2768      3  %  36.2768          3
I                          126     30.5489      4  %  30.5489          4
A                          112     26.7303      5  %  26.7303          5
TO                          95     22.6730      6  %  22.6730          6
```

3.2 カナまたは英文字，分かち書き，紙テープ入力

〔入力例〕　001セイ　マズ　ゴ　シユジン　ヲ　ナント　オヨビ　シ　タラ　ヨロシイン　デシヨﾟ　ウ　。

002カオ　（　オカミ　サン　）ドン　ネエ　，コノ　ゴロ　ハ　ネエ。

先頭から頁行（または識別コード，この例では会話番号と話手コード）5文字を入れる。データは9字目から，3000字以内。それを越える場合は，ギャップを入れてから打つ。頁行はそのブロックで全レコードにつく。

〔使用プログラム〕　紙テープ読み込み (NINPUTPT)→コピー (DUP)→かな用例つけ (NYOUREI 0)→50音順ソート (SORT)→ラインプリンタ出力 (NOUTPUT 1)

紙テープ読み込み (NINPUTPT)→マージカウント (NCOUNT)→度数順ソート (SORT)→比率計算 (KEISAN)→全体度数順表作成 (NPRINT)

〔出力例〕

```
ﾐﾀﾞｼ        ﾍﾟｰｼﾞ,ｷﾞｮｳ                    ﾖｳｼｲ                              PAGE-0001
,           002,ﾏｻ ﾃﾞ ｼﾎﾀﾞﾊ ｸﾃﾙ ｴｰ ｱﾂﾞﾒﾅｰ  ,  ﾝﾀﾞｰ ﾝﾀﾞｽﾞﾅｰ ,
            001,ｷｴ  ﾏｽﾞﾊｰ ﾒﾂﾞﾀｴ ﾐﾗﾝﾈﾅｰ ﾝｰ  ,  ﾐﾝﾅ ﾊﾝﾀﾞｴﾃﾞ ｸﾂﾞﾀﾂﾞﾃ ﾄｼﾖﾛﾘﾀﾞﾊﾞﾘ
            003,ｷｴ  ｱﾂﾞﾒﾅｰ , ﾝﾀﾞｰ ﾝﾀﾞｽﾞﾅｰ ,
ｱﾂﾞﾒﾅ        002,ﾏｻ  ｵｼﾞｴﾝﾃﾞ ｼﾎﾀﾞﾊ ｸﾃﾙ ｴｰ  ｱﾂﾞﾒﾅｰ  , ﾝﾀﾞｰ ﾝﾀﾞｽﾞﾅｰ ,
ｴｰ          002,ﾏｻ ﾂﾞﾆ ｵｼﾞｴﾝﾃﾞ ｼﾎﾀﾞﾊ ｸﾃﾙ  ｴｰ  ｱﾂﾞﾒﾅｰ , ﾝﾀﾞｰ ﾝﾀﾞｽﾞﾅｰ ,
ｵｼﾞｴﾝﾃﾞ     001,ｷｴ                          ｵｼﾞｴﾝﾃﾞ  ｸﾃｿﾅﾅ ﾏｽﾞﾊｰ ﾒﾂﾞﾀｴ ﾐﾗﾝﾈﾅ

ﾐﾀﾞｼ                ｼﾞﾖｳﾎｳ ﾄﾞｽｳ      ﾋﾘﾂ        ｼﾞｭﾝｲ      ﾌﾟﾌﾟﾝﾋﾘﾂ  ﾌﾟﾌﾟﾝｼﾞｭﾝｲ

,                            3    136.3636      1  % 136.3636          1
ｵｼﾞｴﾝﾃﾞ                      2     90.9090      2  %  90.9090          2
ﾝﾀﾞｰ                         1     45.4545      3  %  45.4545          3
ﾝｰ                           1     45.4545      4  %  45.4545          4
ﾒﾂﾞﾀｴ                        1     45.4545      5  %  45.4545          5
```

3.3 漢字かなまじり，分かち書き，読みがなつき，紙テープ入力

〔入力例〕　(P 001・L 01) C/R 親譲り〔おやゆづり〕C/R の C/R 無〔む〕C/R 鉄砲〔てっぽう〕C/R で C/R …

先頭（で囲まれた頁行情報を PXXX・LXX の形で入れる。本文は分かち書き部分を C/R で切る。よみがなは〔 〕に入れて語形の後に入れる。題は@ C/R をその先頭に入れる。段落は＃C/R で示す。行は (LXX) C/R で示す。

〔使用プログラム〕　フローチャート，2—1，2—2，2—3，3の通り。

エラー出力無視

〔出力例〕

「歌謡曲」

彼	079 24	あの時に天皇陛下万歳と三度叫んだ	彼　の声をその儘書いて送らうか涙て書いた
ア	063 14	ノネと答へる山の木霊の優しさよ「	ア　ノネ」「何にさ」彼は言へない二人は若
	063 12	風は貧困二人は若いアノネと呼べば	ア　ノネと答へる山の木霊の優しさよ「アノ

見出し	情報	度数	比率	順位	部分比率	部分順位
の	11	649	73.557	1 ●	73.557	1
に	11	270	30.601	2 ●	30.601	2
は	11	261	29.581	3 ●	29.581	3

3.4 漢字かなまじり，単位・よみがな・語種・品詞・活用情報つき，紙テープ入力

〔入力例〕　(P 00 1・L 0 1) C/R C 親譲り〔おやゆづり〕(S1) C/R L の (WR) C/R C 無〔む〕(T6) C/R S 鉄砲〔てっぽう〕(T1) C/R L で (WR) C/R…

先頭（ ）内に頁行情報を PXXX・LXX の形で入れる。本文は単位切りし，単位情報とC/Rで囲む。単位情報は三種類入れることができる。最も長い単位の切れ目にCを，次に長い単位の切れ目にLを，最も短い単位の切れ目にSを入れる。よみがなは〔 〕に入れて語の後に，語種・品詞・活用情報はその順にコードで（ ）に入れてよみがなの後に書く。題は@ C/Rをその先頭に入れる。段落は＃C/R で示す。行は(LXX)C/Rで示す。

〔使用プログラム〕　フローチャート2—1，2—2，2—3，3の通り。

エラーはエラー修正ルーチンで直す。

〔出力例〕

```
                361  01  つた木立の辺に来たので，男子王は  足  を休めた．「ねえさん，ここらで初るの
脚              341  01  めので，とうとう赤松の幹のやうな  脚  に縋つた．「姉さん，これはどうした
足音            332  10  と，此材木の間へ人の這入つて来る  足音  がした．「はて何かい」と母親が声を掛
足踏            366  10  くの間ひつそりとしてゐる．三郎は  足踏  をして，同じ事を二三度繰り返した．
あす             356  04  水が温み，草が萌える頃になつた．  あす  からは外の農事が始まると云ふ日に，
                346  06  うとするので，一度は吃られたが，  あす  からは路々が賑ひに来ると言つて，や
```

見出し	情報	度数	比率	順位		部分比率	部分順位
て	WP	112	7.646	15		0.000	0
ゐる	SEGF	96	6.553	16	●	14.111	1
へ	WR	93	6.348	17		0.000	0
から	WR	78	5.324	18	●	0.000	0

4. あとがき

電子計算機によって作られる文脈つき用語索引が言語研究に有用であることは言うまでもない。特に，漢字プリンタの利用は国語研究者をますます計算機に近づけるだろう。これまで，数年あるいは数十年がかりでつくられた索引類も計算機利用によって簡単に作ることができ，研究者を真の国語研究に打ち込ませるだろう。金田一春彦氏らの「平家物語総索引」はそのあらわれである。

索引作成のプログラムはもっと簡単にならなければならない。簡単の意味は二通りある。一つは誰でも簡単に計算機をつかい，簡単に操作ができることである。これは，計算機の普及とオペレーティングシステムの簡略化にまたねばならない。一つは，自動処理である。原文をそのまま入力するだけで，望みの索引なり，集計表を得るシステムである。これは，現在「一貫処理の研究」として開発中であり，ほぼ90％の正解率を持つシステムが完成している。報告は別の機会にしたい。

計算機利用が万能でないのは，これまた言うまでもない。その最大の欠点は，従来のカード整理の途中にしばしば生まれたひらめきが少なくなったことである。特に索引作成のように，それが国語研究そのものでなく，手段である場合によく言われる。しかし，これは計算機利用の目的と矛盾するものである。機械と人間の最適合システムは現在最も重要な問題である。村木新次郎氏の用例のカード形式出力や斎藤秀紀氏のターンアラウンドシステムはこの問題

の一つの解決法を示していると思われる。

このような利点や欠点を持ちながらも，この「索引作成のためのプログラムライブラリ」は我々の仕事の機械的部分を肩がわりしてくれることと信ずる。

このプログラムシステムの作成・整理・実行には研究補助員長田厚子嬢，アルバイタ稲垣雄次・片岡利徳両君の協力があった。記して感謝の意を表する。

参考文献

石綿敏雄（1971）　新聞用語調査の用例印字プログラム　"COBOL-KWIC"（国研報告39「電子計算機による国語研究Ⅲ」）

植村俊亮（1975）　電子計算機による自動索引の研究（上下），（電子技術総合研究所報告第734号，第747号）

金田一春彦・清水功・近藤政美（1973）「平家物語総索引」（学習研究社）

斎藤秀紀（1968）　電子計算機と漢テレによる用語総索引の作成。（国研報告31「電子計算機による国語研究」）

坂本義行・岡本哲也（1975）　日本語のコンコーダンス。（情報処理学会計算言語学（ＣＬ）研究資料２）

坂本義行（1976）　日本語文献テキストの自動処理について。（情報管理18—10）

田中章夫（1468）　電子計算機によるワードリスト作成上の一問題。（国研報告31「電子計算機による国語研究」）

土屋信一（1972）　カナ入力による日本語文総索引の作成。（国研報告46「電子計算機による国語研究Ⅳ」）

鷲岡昭夫（1973）　文語形・口語形活用語の代表形の変換処理について。（国研報告49「電子計算機による国語研究Ⅴ」）

（1972）　用語検索システムの言語単位。（国研内部資料ＬＤＰ月報別冊10）

（1976）　言語研究のための索引作成システム。（国研報告「電子計算機による国語研究Ⅷ」）

中野　洋（1972）　用語検索のための付加情報について。（国研内部資料ＬＤＰ月報別冊10）

（1973）　品詞情報つけの規則。（国研内部資料ＬＤＰ月報別冊11）

（1973）　入力データのチェックシステムについて。（国研内部資料ＬＤＰ月報別冊11）

（1976）　語彙調査の一貫処理システム（情報処理学会計算言語学（ＣＬ）

研究会資料５）
中野洋・鷺岡昭夫（1974）　漱石鷗外の用語索引の作成。（情報処理学会ＣＬ研究委員会資料1974-4）

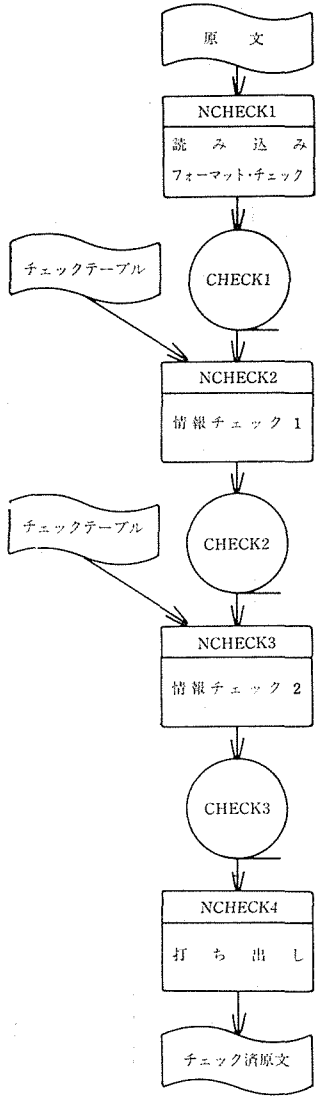
プログラム一覧，説明ページ

番号	プログラム名称	PROGRAM-ID	ステップ数	ページ
101	データの読み込み	NCHECK 1	403(255)	24
102	情報チェック１	NCHECK 2	350(179)	26
103	情報チェック２	NCHECK 3	270(140)	28
104	紙テープ打ち出し	NCHECK 4	147(93)	29
201	紙テープ(原文)読み込み	NINPUTSKN	419(269)	31
202	フォーマット変換	NDATASAKU	622(400)	32
203	活用形換変	NKATSUYO	904(718)	34
204	配列情報つけ	NGOJYUON	457(304)	35
205	コピー	DUP	——	35
206	かな用例つけ	NYOUREIO	193(93)	35
207	漢字かなまじり用例つけ	NYOUREI 1	208(105)	36
208	併合	NMERGESKN	146(75)	36
209	50音順ソート	SORT	——	36
210	カード(原文)読み込み	NINPUTCR	115(55)	37
211	紙テープ(かな原文)読み込み	NINPUTPT	150(85)	38
212	Ｌ・Ｃ単位作成	NLCTANI	219(116)	38
213	ラインプリンカ出力	NOUTPUT1	153(62)	39
214	漢プリ出力用編集	NOUTPUT2	346(222)	39
215	漢テレ出力	NOUTPUT3	183(66)	40
216	原文出力	NTEXTSKN	240(121)	40
217	漢プリ出力	CVT-MT, MT-HKP	——	41
218	データチェック	NCHECKMT	740(559)	41
219	紙テープ打ち出し	NCHECKOUT	147(93)	42
220	修正パラメータ作成	NPARAM1	498(306)	42
221	データ修正	NSYUSEIMT	115(67)	43
222	フォーマット変更	NKATAGAE	327(209)	44
223	一語検索	NGOR	53(16)	44
224	一品詞検索	NHINSHIR	53(21)	44
301	50音順ソート	SORT	——	46

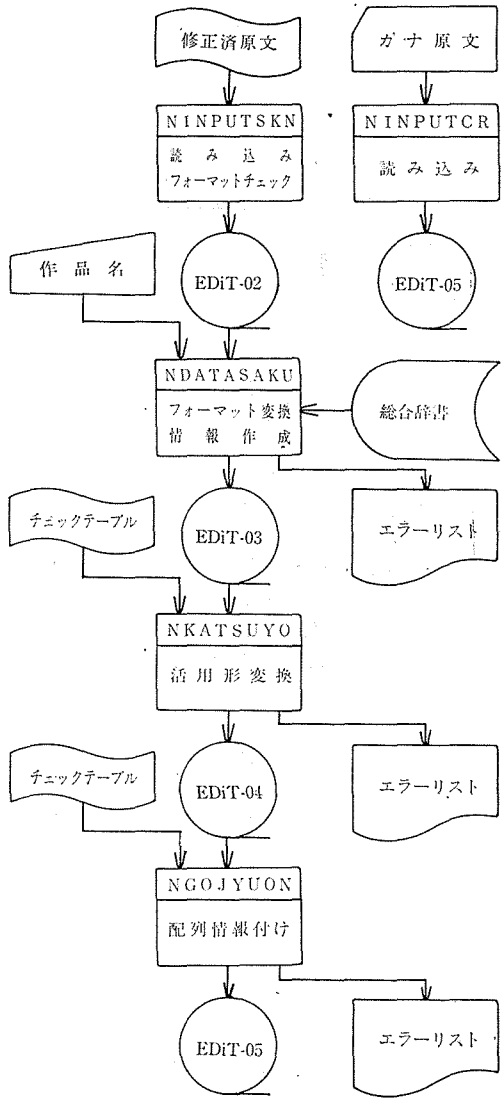
302	マージ・カウント	NCOUNT	147(78)	47
303	度数順ソート	SORT	——	47
304	比率計算	NKEISAN	86(32)	47
305	全体度数順表作成	NPRINT	117(40)	48
306	語種別集計表作成	NSYUKEI 1	117(40)	48
307	品詞別集計表作成	NSYUUKEI 2	117(40)	48
308	抜き出し	NPICKUP	111(54)	49
309	漢プリ出力用編集	NKLPPINT	403(190)	49
313	漢プリ出力	CVT-MT, MT-HKP	——	49

注) PROGRAM-ID の SORT, DUP, CVT-MT, MT-HKP はサービスルーチンである。それ以外はすべて COBOL で書かれている。( ) 内の数字は PROCEDURE DIVISION のステップ数である。

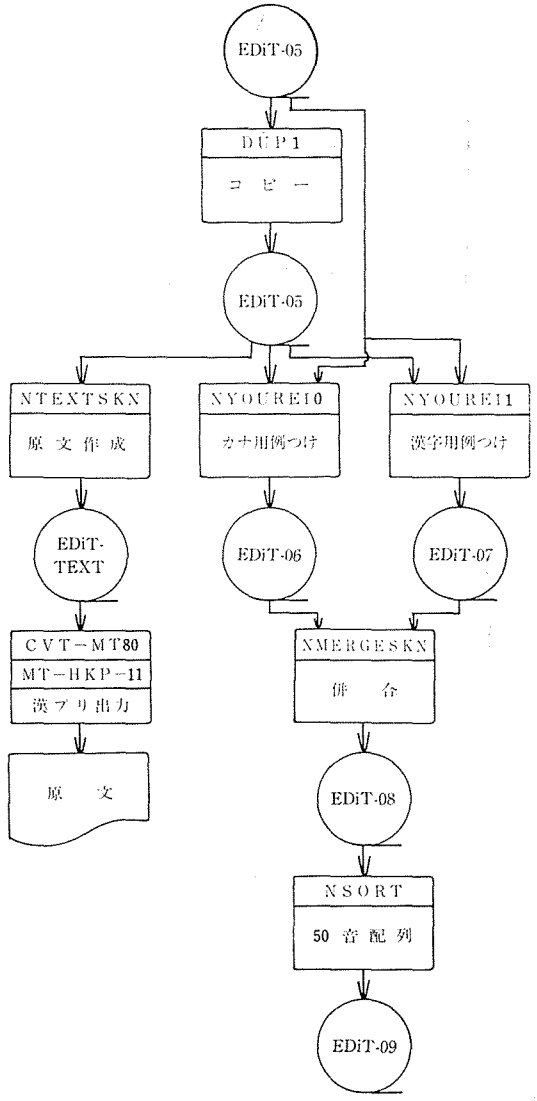
フローチャート　1　チェックルーチン



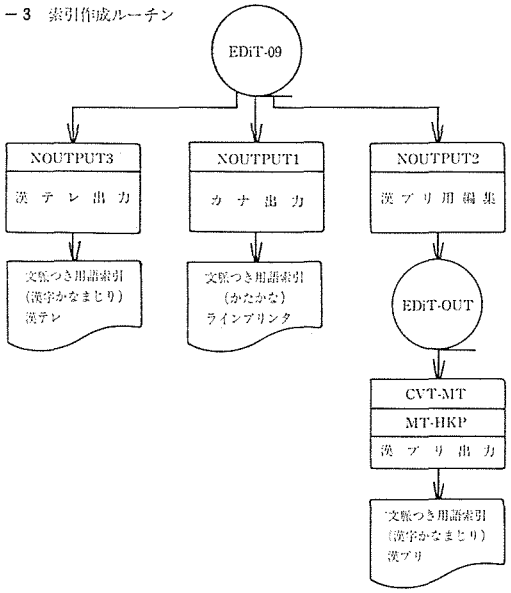
フローチャート２－１　索引作成ルーチン



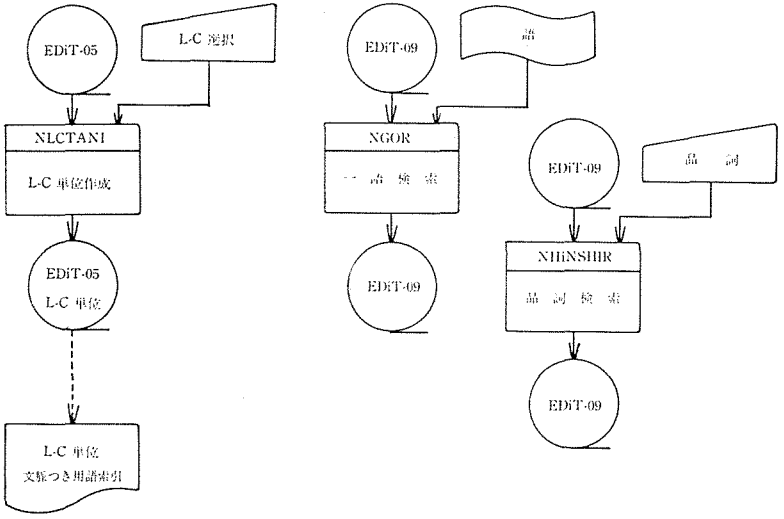
フローチャート　2－2　索引作成ルーチン



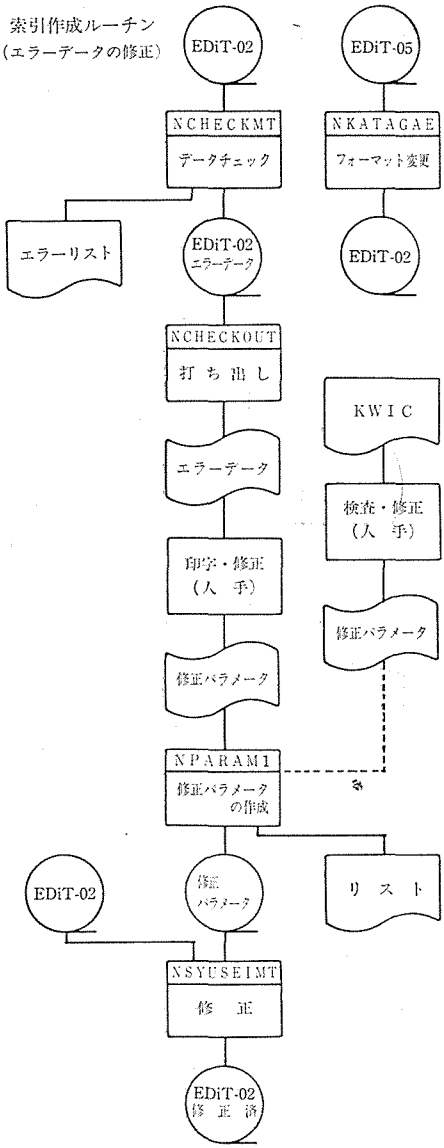
フローチャート　2－3　索引作成ルーチン



（特殊処理）



フローチャート　2－4 索引作成ルーチン
（エラーデータの修正）



フローチャート　3 ワードカウントルーチン