カナ入力による日本語文総索引の作成

著者	土屋 信一
雑誌名	電子計算機による国語研究
巻	4
ページ	35-43
発行年	1972-03
シリーズ	国立国語研究所報告 ; 46
URL	http://doi.org/10.15084/00001012

カナ入力による日本語文総索引の作成

土　屋　信　一

1. はじめに

これまで，国語国文学研究のために，いくつかの文献の語彙索引が作られてきたが，それらは皆，ぼう大な年月と人力とによるものであった。それゆえ，研究者にとっては，索引作り それだけが一つの大事業と 考えられてきたし，「ためしに」索引を作ろうという安易な気持ちから索引作りを試みる人もいなかった。ために，テキストに問題のある「枕草子」の索引はなかなか作られなかったし，テキストに問題がありかつ大部である「宇津保物語」「平家物語」などの索引はいまだに作られない状態である。

そこで，計算機を使って，少しでも人手を省くことができないかという考えのもとに，索引作成システムを企画した。わずかでも人手が省かれた結果，研究者が索引作りから，それだけ解放されるとか，気安く索引を作り，それを手段として研究をすすめるという方法が取られる，とかいう傾向が少しでも現われれば，この企画の目的は達成されたことになる。

2. システム設計のねらい

システム設計に際して，必要な条件を個条書きにする。

(1) 汎用性があること。(一つのシステムで，どの古典の索引も作れること。索引の用途に応じた単位が選べること。)

(2) できるだけ安い経費で上げること。

(3) 索引は印刷物としてアウトプットされること。(利用者は計算機を使わずに索引を利用し得ること。)

(2)のためには，人手による方が安い作業（例：単位切り）は人手を使うこと，計算機の使用時間をできるだけ短くするためシステムを単純にすること，

などが考えられる。

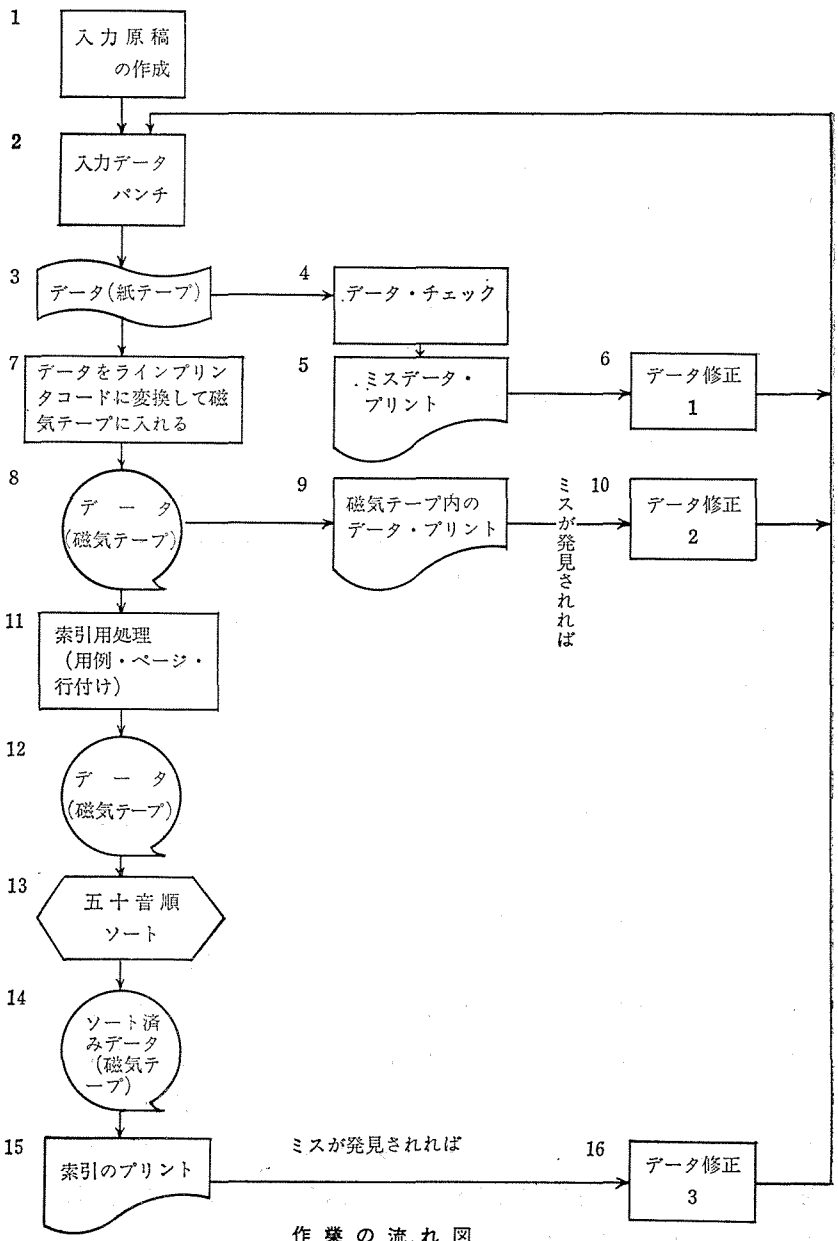
また，これは重要な条件ではないが，システム設計とプログラミングを私一人でやらなければならず，また，それにあまり時間をかけたくないので，あまり複雑なシステムを設計しない方が望ましい。そのため，新しい情報処理の方法の導入とか開発は全く考えなかった。なお，私は，電子計算機というものは，ＨＩＴＡＣ—3010しか知らないので，ここで考える機械処理も，すべて3010の能力の範囲内にとどまる。

3. システムの概要

最初に入出力機器について述べる。入力装置はフレキソを用いることにした。漢テレを用いなかったのは，出力の際多くの時間を要するのと，どこにでもある機械ではないという理由からである。フレキソならば，入力作業を外注することも可能である。出力装置はラインプリンターを用いることにした。フレキソを出力装置として使うことも可能であり，その方が便利な点もあるが，機械のスピードと安定性から，ラインプリンターを選んだ。

入出力装置が決まった結果，索引に使う文字はカタカナ・ローマ字・数字・その他の記号の中から選ぶという制限ができた。本文には，この中からカタカナを選ぶことにした。（ローマ字を選ばなかったのは，日本語の索引には，カタカナの方が読みやすく，また便利であろうという判断による。）使用する文字・記号は，カタカナ48，長音符号，句読点，濁点，半濁点，かっこ（　）の5種類である。この55種類があれば日本語の音節はだいたい表わすことができる。半濁点をさまざまな用途に使うのも一方法である。たとえば，万葉仮名の甲乙も，乙類に半濁点を付するという方法を取れば表わせるし，「オトッツァン」の「ツァ」も「サ°」と表わすことができる。ただ，拗音・入声音などで普通小書きにされる文字ア・イ・ウ・エ・オ・ツ・ヤ・ユ・ヨを半濁点で表わすのは，配列の際やや不便である。また，かっこによって，さまざまの情報を付加することも可能である。

次に，作業の流れについて述べる。次に掲げたのだが，その流れ図である。入力作業〔(1)・(2)〕およびデータ修正作業〔(6)・(10)・(16)〕が人手による作業


1
入力原稿の作成
2
入力データパンチ
3
データ(紙テープ)
4
データ・チェック
5
ミスデータ・プリント
6
データ修正1
7
データをラインプリンタコードに変換して磁気テープに入れる
8
データ(磁気テープ)
9
磁気テープ内のデータ・プリント
ミスが発見されれば
10
データ修正2
11
索引用処理(用例・ページ・行付け)
12
データ(磁気テープ)
13
五十音順ソート
14
ソート済みデータ(磁気テープ)
15
索引のプリント
ミスが発見されれば
16
データ修正3

作業の流れ図

で，他はすべて機械による作業である。機械処理については次節に述べる。ここでは人手による作業のあらましを述べよう。

○入力作業〔(1)・(2)〕

(1)はテキストを単位に切り，それをカタカナに改め情報を付しながら入力原稿を作成する作業である。この間に何度かの校合が行なわれる。入力データは1行を1レコードとする。原稿の形式は次のとおりである。

・ページ……3けた（pのあとにしるす）

・行　……2けた（Lのあとにしるす）

・本　文……1行最大200字（濁点・半濁点・長音符号は1字，単位切りのため挿入したスペースも1字と数え，かっこ・句読点は3字，ヰ・ヱ・ヲはそれぞれ4字と数える）

◦1行が200字を越えるときは，行を改めれば入力することができる。

◦本文の前には必ずＬＣ（オクタル57），おわりにはＵＣ（オクタル17）とＣＲＬＦ（オクタル55）がなければならない。

◦句読点およびがっこの前にはＵＣ，あとにはＬＣが必ずなければならない。

◦ヰ・ヱ・ヲの前にはＭＣ（オクタル37）あとにはＬＣが必ずなければならない。

・単　位……1単位は最大15字，句読点・かっこを付けても最大40字とする。

○データ修正作業〔(6)・(10)・(16)〕

データのミスの検出と修正は3個所で行なう。（このほか，(1)(2)の段階でも，前述のようにミスの検出が行なわれる。）このうち，計算機によって検出したミスを修正する(6)が，作業の中心であり，(10)は(6)で発見し損ったミスを検出するためのもので，あまり検出能力はない。また，(16)では同語形が集められ，一応索引の形式になっているため，単位切りのミスなど(1)の段階で起きた人間的な（高度な）ミスを検出することができる。以上のようにして検出されたミスの修正はすべて入力原稿の段階で行なうことにした。

4. プログラム

このシステムにおけるプログラムは次のとおりである。かっこ内の番号は流れ図の該当番号である。

Aデータ・チェック (4)・(5)

ここでは主として(1)の原稿作成の段階における勘違いによるミス，および(2)のパンチ・ミスを検出するのが目的である。ミス・データがある場合，その所在とミスの内容とがラインプリンターによって打ち出される。検出するミスの種類は，レコードのフォーマットに関するもの２種，ＵＣ・ＬＣ・ＭＣなどのキーの使い方に関するもの７種，かっこの使い方に関するもの１種，レコード・単位の長さに関するもの３種の計13種である。

Bコード変換 (7)

ここでは，フレキソ・コードで書かれたレコードをテーブルを用いてラインプリンター・コードに変換するのがおもな仕事である。その際，ＬＣは＝（イコール）に，ＵＣは’（アポストロフィ），ＣＲＬＦはＥ/Ｉに変換される。その結果，96文字モードでプリントすると，ＬＣとＵＣのあいだのすべての文字はカタカナ・濁音符・半濁音符・長音符で表わされる。なお，ヰ・ヱ・ヲは，それぞれWイ・Wエ・Wオで表わされる。また，このランでは，各レコードのはじめにページと行とを付す作業もする。（原稿では，ページのはじめのみページ数をしるしてある。）

C磁気テープ・データ・プリント (9)

これは，磁気テープに収められたデータを96文字モードで打ち出す作業である。その際，レコードの総数も計算し，結果をプリントする。この数字は全データの行数の合計と一致するが，それよりも少なければ，数行が１レコードとなっている（レコード間にギャップがはいっていない）可能性があるので，データの検査をする必要がある。

D索引用処理 (11)

ここでは，スペースによって切られた１単位を見出し語とし，その語を中心におき前後90字（語頭からうしろ47文字，前の語の最終の文字から前43字）を用例として付し，最後にその語の存在するページ・行を付したレコードを作成

する。用例を付する作業において，メモリー内には，その語の属する行と前後の各1行とがおさめられており，その中から90字が取られる。そのため，1行を構成する字数が小さい場合，90字分用例が付けられないこともありうる。このことは，このプログラムが短い行のデータを扱うのに適していないことを示している。一行平均45字以下のデータを扱う場合は，この部分のプログラムを差しかえる必要があろう。

なお，1単位を見出し語として立てる際，かっこおよび句読点は除いた。これは同語形の語を集める際その方が便利だと考えたためである。この結果，句読点の数を数えることや，文末の語形のみを取り出すことができなくなった，などの欠点も生じた。

E五十音順ソート (13)

これは，見出し語を五十音順に配列し，同語形の中ではページ行数順に配列する作業である。サービス・ルーチンを使用するため，技術的には容易である。ただ，同ページ同行の同語形は，その出現順に配列することができないため，テキストの順序と異なることがある。また，見出し語を語尾から五十音順に配列することも容易にできる。

F索引プリント (15)

これは，五十音順配列の済んだデータを索引の形で打ち出す作業である。一応1ページ50語ずつプリントすることにした。また，このプログラムは，(12)のデータもプリントできるし，(13)で五十音順以外の配列を行なった場合もプリントできる。

5. 今後の問題

以上で，このシステムの概要について述べたが，今後，この索引作成システムをよりよいものにしていくために，すでに気のついているいくつかの事項についてしるす。

(1) 入力の際の情報を入れやすくすること。

現在，入力原稿はカタカナに限られ，（ ）によって付せられる情報もカタカナに限られている。そのため，表記情報を付けるのが不便なので，改良しな

ければならない。現在でも，原文のかなづかいを示したり，「行なう」を「オコナウ（ーーナウ）」と書いて送りがな表記を示すことはできるが，さらに漢字を表わす工夫が必要と思われる。また，アルファベットも使えるようにし，品詞その他の情報を付けようと思えば付けられるようにする必要もあろう。「　」・？・！なども必要と思われる。

(2) 用例をもっと長くすること。

現在，90字の用例を付しているが，これはラインプリンターの字数から割り出したものである。さらに用例を長くし，カード1枚分程度にして磁気テープに収めておき，必要度に応じて適当な長さに印字するようにするならば，これまでのカードによる検索システムに完全に代わり得るであろう。

(3) 単位の切り方を数通りにすること。

現在，単位の切り方は，一通りだけであるため，いく通りにも切れるもの，または，切った方がよいと思われるものも，一通りにしか切ることができない。そのため，掛けことば・地口・しゃれの類の処理が不十分なので，これを解決するため，いく通りにも切れるようなシステムを考えたほうがよいと思われる。

(4) このシステムをだれでも利用できるようにすること。

このシステムのプログラムは，すべてアセンブラ言語で書かれている。そのため，ＨＩＴＡＣ3010を使う人には便利であろうが，他の機種を使う人には全く役に立たない。今後システムを改良すると同時に，プログラムもコボルなどのコンパイラに書き改めることが望まれる。

(5) 語と語形と表記との関係について研究すること。

このシステムでは，テキストの語形がそのまま見出し語として表わされる。そのため，語形をどのように表記しておくかが重要な問題となる。しかし，語と語形と表記との関係についての研究がまだ不十分であるため，語形表記の方法が確立していない。

たとえば，具体的には，「こりゃ大変だ」という文を入力するさい，「コリャタイヘンダ」と書き改めたのち，どう単位を切ればよいだろうか，という問題が起こる。「コリャ」を1単位とするのは不便である。だからといって，

「コリ／ャ」とするなら，「コリ」「ヤ」という表記は語形を表わしているか疑問であるし，また，検索する側からいっても，「コリヤ……」を見つけ出すのに「コリ」「ヤ」を使うのは容易ではない。語は「これ」「は」であり，その表記は「コレ」「ハ」であると考え，「コリ」「ヤ」はその表現上変化したものと考え，入力では，「コレ（コリ）／ハ（ヤ）」または「コレ（コリヤ）／ハ（コリヤ）」とするのも一方法であろうし，入力にローマ字を使って，「ＫＯＲＥ（ＫＯＲ）／ｗａ（ｙａ）」とするのも一方法であろう。

これは一例であって，語と語形との関係の方が重要なことがらと思われるが，まだ十分に考えていない。語と語形との関係を研究することによって，ここに作成された索引が，「語形索引」から「語彙索引」に近づいてゆくことが予想され，また，これまでの「語彙索引」についても考え直さなければいけないことがらが起こってくることが予想される。

追記：この索引作成システムの実用第１号として，洒落本「遊子方言」（岩波古典文学大系59「黄表紙洒落本集」ｐ270～ｐ294）の索引を作成した。その一部を，最後に掲げる。

アサ	ムスコ W オ ツレ , フタリ ニハ ヘ オリル . ニヨウボウ ミヨウ	284.16
アサギ	ヨウボウ ハイ , トオリモノ テノコヴイ , オレ カ ヤロ . ト ,	279.10
アサツテ	. モウ ヨシワラ モ , フタタビ ナイ ワイ ナイ ワイ . ムスコ マタ	291.14
	デ ナサ ンセ ン カ イ . トオリモノ ナセ . ムスコ ワタシ ヤ ,	291.15
アシ	″ テ ル , ヒヤウシ ニ , トオリモノ ノ アシ ニ , ワカイ モノ ノ	282.12
	, カマハ ズ , トン テ テ ル , ヒヤウシ ニ , トオリモノ ノ	282.12
	キヤク ニ ムカヒ テ , オ イテ アソハシ マシ タ カ . スツキト オ	289.10
アシヴ	ラレ タ . イツツ ノ コト ネ ヨウ ソヴ . アア ヨイ モ サメル .	290.03
アス	リル . ニヨウボウ ミヨウ アサ , オ ヨリ ナンシ . トオリモノ アイ	285.01
アソハシ	モ イハ ズ″ , キヤクシヴン ヒラ ニ ムカヒ , ワカイ モノ オ イテ	282.13
	ノミ ナンス ノ . ト イウ テ , キヤク ニ ムカヒ テ , オ イテ	289.09
アソビ	シヤウス , イハンヤ コノ ハナ ニ オイテ W オヤ . ホクシウ ノ	270.07
アソフ	ハ シツホ°リト , ラマエ ト , オレ ト , タツタ フタリ デ ,	280.12
アタタカ	″イノモノ , カシ , チ゛ユウハ゛コ ノ フタ W オトレ ハ″ ,	289.13
アタツ	トンタ゛ イキ ナ キレ タ゛ . コハセ゛ ハ″カリ , ニフ゛ ホト゛ ニ	276.13
アタマ	キ″ハ″シ テ゛ サンシ゛ユウシコ゛ ノ W オトコ (オトコ) . スコシ	272.02
	サ゛トウ テ゛モ , マタ゛ カラス カ゛ ナカ ン . シンゾ″ウ サ゛トウ ノ	291.07
アタラシキ	オホキク , ヒトカ゛ラ ヨク , アハセ゛ン ニテ , クロ ハフ゛タヘ ノ	282.11
アタリ	カケ . クレ ヨリ ホカ ニ イロ W オトコ ハ ナシ ト . カウマン ニ	272.07
アタ゛ツイ	イル ト , サル ウチ ノ オクサ゛シキ ノ シ゛ヨロウ カ゛ トウシ゛ユウ ニ	278.05
アタ゛ツキ	オクサ゛シキ ノ シ゛ヨロウ カ゛ トウシ゛ユウ ニ アタ゛ツイ タ . オレ ニ	278.05
アチラ	マウシ オ マウシ . カフ゛ロ アニ サン , ソノ シヤミセン ハ″コ ,	289.15
アチ゛	ナル ナカ ニ , ケシカラヌ コ W エ ノ アンマ ハリ , スシ ウリ カ゛	287.10
	オサ゛ンセ ン . ヒラ ハテ キ ノ オオイ ヤツ ノ . ト ツメリ テ ,	293.10
アツ	W オ , シツ タ カオ シ テ イツ タ カ゛ , オラ , ツイト゛ ,	274.14
	ユ カ゛ , オレ ト ココロヤスイ ニ ヨツ テ , イツ デ モ テイシユ ニ	276.15
	″ ウチ ヘ キ サツシヤイ . ツツチ ヘ テカ゛ミ カ゛ ヤリ タイ コト カ゛	277.15
	デ , イマイマシ . ト゛コ ニ シ ヨウ ナ . ソレ ソレ コノ コ゛ロ	279.01
	ヘ イテ゛ , タ゛イフ゛ オ オソウ コ゛サ゛リ マシ タ . ヒラ ソウ テ	283.05
	, ノミ カケ ヤマ ト シ テ イ タラ ハ″ , テイシユ ノ ミエ ラレ テ	283.06
	. モツトモ コレ ハ , トラノモン ノ ホウ ヘ ヨル ト イウ コト デ	283.10
アツイ	ン ト モ コ゛サ゛リ マセ ン . トオリモノ テノコヴイ W オ , チヨト	279.09
アツタラ	ルク スル ト , カナワ ナイ コト カ゛ アル . ソレ ニ ツケ テモ ,	276.04
アツラエ	ヨイ タ゛ . セツニ カ゛ ナイ ト イエ ハ″ , オレ カ゛ マルカ゛ク イ	276.12
	ルエ ツク ホト゛ ナ キレ タ゛ . オレ カ゛ セツニ ノ アル トキ ニ ,	276.15
アツラエル	オレ ト ココロヤスイ ニ ヨツ テ , イツ デ モ テイシユ ニ アツ テ ,	276.16
アテト゛	シ ト . カウマン ニ アタリ W オ . キロキロ ト ミ マハシ テ .	272.07
アト	″イ サ . アリヤ , カマチヤ モトシ゛ロウ ト イウ シ゛シ゛イ タ゛ ,	274.12
	ツ シ゛ヤ ナイ カ ノ . セントウウ シユビ゛ ノ マツ ハ , ヨツホ°ト゛	275.14
	ウチン ミハリ ニテ , ハシコ゛ W オ アカ゛ル . シンゾ″ウ キヤク ノ	288.02
	ウシ ンシ タ ナレ ハ″ , オマエ ノ シヨウヨウ ニ イキ ナンス トキ ,	292.13
	. ト イヒ ナカ゛ラ , ロウカ ヘ オクリ イツ゛ル . ムスコ ヘヤモチ	292.16
アナタ	. カエツ (カヘツ) テ , イオ (W オ) ウ ニ ハ ,	273.07
	. ヤマモトヤ ニヨウボウ ヨウ オ イテ ナサ レ マシ タ . ヤマモトヤ	279.07
	マエ サマ ハ , オ ミ ウスレ モウシイ ンシ テ コ゛サ゛リ マス カ゛ ,	281.05
	ソレテ゛ ミ ナイ カオ シ タ ソウナ . ニヨウボウ IIエ ユウヘ ,	281.12
アニ	, クチクチ ニ , シンゾ″ウ コ゛テ サン , オ イテ゛ ナンシ .	289.06
	ナニ イマ ノ サカス゛キ ハ , コウ オ マウシ オ マウシ . カフ゛ロ	289.15
アノ	ホウ ヘ イク カラ , ヒラ サン ト カウ サン カ゛ コ゛サ゛リ タラ ,	282.06
	. セント゛ オマエ ノ ヨイ ナンシ タ トコロ テ゛ オサ゛ンス . キヤク	288.05