An automatic processing system of natural language

言語処理における一貫処理法の研究

中野　洋

1. はじめに

電子計算機を利用して生ずる最大の利点は，それを使わない場合に比べ，はるかに人的作業が少なくてすむことである。機械的な作業が少々多く，時間がかかっても，機械は24時間働くことができるからいい。人間はそうはいかない。

ところで，言語処理——特に，語彙調査や索引づくりの場合，集計・配列作業は計算機むきで簡単に実行できるが，言語的な情報の付加作業は複雑で，従来はこれを人手の作業にたよってきた。しかも，これらの作業は，時間も費用も労力も膨大な量を投入しなければならないのが現状であった。それでも処理量が非常に多い場合には計算機むきの仕事と人間むきの仕事を分離することができ，計算機を使うメリットもうまれるのだが，少量の処理ではかえってすべて人手でやった方が能率的だという場合が起こる。これでは計算機本来の利点が損われてしまうことになる。

そこで，現在人間がおこなっている複雑な作業——言語的な情報の付加作業を計算機に肩がわりさせ，人的作業を軽減するシステムを考えた。これを我々は一貫処理システムとよんでいる。

人的作業の種類——特に電子計算機による語彙調査や索引作りにおいて——

電子計算機による言語処理，特に日本語の処理において避けることのできない過程の一つに，入力文の単語分割（語彙調査などにおいては，調査単位による分割という意味で単位切りと呼んでいる）がある。これは，語彙調査に限らず，機械翻訳にしても，情報検索にしても，自然語文を処理する場合には避け

ることができない。また，処理結果を単語レベルで出力する場合，人間が見なれている順序——たとえば50音順に出力するために漢字にはよみがなをつける必要があるし，語彙研究のためには，語種・品詞別の結果を得るために語種・品詞情報を付けておきたい。また，文・文章レベルの処理分析においては，各種の文法情報が必要になろう。

以上は，人間においてもかなり能力を要する作業だが，データをマシン・リーダブルにするための作業，すなわち，データの清書・パンチ・校正などは，光学文字読み取り機械・音声認識機械の実現をみない現状では，避けることのできない，時間と費用とがかかる作業である。

これらの他に，その目的によって，特有の，人間でしかできない複雑な知的作業がある。語彙調査においては調査単位の設定や同語異語の判別があり，文献検索では抄録の作成，キーワードの付加，シソーラスの作成などがある。これらの機械化が実用レベルに達するのは少し先のことになろう。

ところで，語彙調査の作業工程の中で，人手による作業は次に示すとおりである。人手でおこなった最大の語彙調査である雑誌九十種の調査を例に示す。そのうち，破線（……）で示したのは，電子計算機によって語彙調査をおこなっても残った手作業であり，実線（——）で示したのは，一貫処理法によってもなお残る人手の作業である。それ以外のすべては，方法はかわるが機械によって処理可能である。

〔語彙調査作業工程〕

1．準備

10. 文献入手→文献入手→文献入手

11. 標本抽出→サンプリング割り当て・補正→サンプリング割り当て・補正

12. 採集用カード作成→清書・データパンチ・校正・修正パンチ

2．採集

21. 単位語に分割・指定→単位語に分割

22. カード採集

23. 22の検査→校正・検査・修正パンチ

3．整理

30. 集落ごとに，検査済み採集カードの枚数確認

31. 集落ごとに，採集カードの五十音順排列→このための，よみがな付加パンチ

32. 31の結果の整理票所定欄への転記→終止形変換のための，活用情報の付加・パンチ

33. 31，32の検査→「よみがな」の検査・修正パンチ

34. 排列の一本化と整理票照合→同語異語の判別→同語異語の判別

35. 派生語等の親票作成

36. 整理票のパンチング

4．集計

41. 延べ語数の確定

42. 使用率計算

43. 精度の計算

5．製表

51. 整理表（子票以外）の使用率順排列

52. 使用率順語彙表作成

53. 整理表（親子共）の五十音順排列

54. 五十音順語彙表作成

いいかえれば，一貫処理システムでは，文献を手に入れること，調査の対象を決めること，同語異語判別をすることの三点だけを人間がやり，あとのすべてを機械がやるというシステムを目指しているのである。

人的作業軽減のねらいと効果

人的作業を軽減すると経費や時間の節約につながる。しかし，ねらいはそれだけでなく，我々の研究の目的がよりよい結果を得ることにあり，かつ語彙調査等の言語処理にともなう作業が膨大でやるべきことを満足にやれないという現状では，まず浮いた経費や時間をより人間的な高度の知的作業にむけることができる。

人間の作業の多くは高度な知的作業だが，同時に簡単なミスをたびたび犯

す。すべてを人間の作業によるのならそれ以降の工程でミスを発見することもできる。しかし，人間の作業結果を機械に処理させた場合，発見は困難となり，ミスはミスのままで処理されてしまう。その結果，処理の精度がおちる。これを避けるには，検査に時間をかけなければならない。一方これを機械にまかせることができれば，作業の程度は低くなるが，人間のミスがどこに現われるかわからないのに対し，機械の処理ミス……人間からみればミスだが，機械にとってはプログラムどおりに作動した当然の結果……は多いけれど，現われ方は一定になる。それだけ，発見が容易で修正もしやすくなる。

ところで，機械にできる人間の簡単な作業（たとえば，清書・フォーマット変換・簡単な単位切り，情報つけ）が少なくなるということは，それだけ人間のミスをおこす機会が少なくなるというわけである。人間の作業の結果はパンチによって入力されるわけだから，機械化によってパンチの作業量も減ることになる。

人間によって起こるミスの影響を少なくするためには，人間の作業を工程の後の方に持っていく方がよい。また，機械によって起こるミスを検査・修正も工程の後に入るわけだから，一貫処理の人間作業は必然的に後に集中することになる。

2. 一貫処理を実現する方法

一貫処理をするためには，次の二点が満足されていなければならない。一つは，大量に蓄積されたデータがあるということ，他の一つは処理のプログラムが用意されていることである。この二点とも，国語研究所は満足しているのであるが，いま少し詳しい説明を試みよう。

2—1 蓄積データの利用

2—1—1 国語研究所外にある言語データの利用 電子計算機による写真植字（電算写植）による印刷は，最近急速に増えつつある。関係者に聞くところによると，1985年には印刷業界の8割は電算写植になるという予想だそうである。ところで，電算写植の中間出力としてマシンリーダブルなデータがある。現在は，紙テープが多いが将来磁気テープや他の媒体になることもあろう。と

にかく，マシンリーダブルな形になっているのだから一貫処理システムに接続することができる。これが実現すれば，入力用のパンチ量が大幅に少なくなる。一貫処理システムはこれを可能にするシステムである。

言語情報処理を業務として，あるいは研究の対象としているところはたくさんある。たとえば，国立国会図書館の文献の索引づくりや日本科学技術情報センタの科学技術情報のサービスなどは，国語研究所の語彙調査などと同じように大量の言語データの作成・処理をおこなっている機関である。これらの機関の言語データを使うことができれば，入力データパンチが少なくなる。

また，言語情報処理を研究の対象としている機関，たとえば，電子技術総合研究所・京都大学工学部・九州大学工学部・武蔵野通信研究所などでは言語処理のためのアルゴリズムの開発とともに，ある程度の実用をねらうために機械処理用の辞書を作成している。この辞書の作成の一つの方法として，既存の国語辞典や英和辞典，英英辞典を入力し利用しようとしている。これらのデータはもちろん一貫処理用に有用なデータである。

2—1—2　国語研究所内にある言語データの利用　　1966年に導入された国語研究所の電子計算機が処理したデータ量は，後に示すように延べ450万語になろうとしている。これらのデータの多くには各種の情報が付けられ磁気テープに納められている。これらを言語処理用の辞書とすることによって少なくとも人的作業やパンチ量を少なくすることができる。

国語研究所のデータ一覧

(ア)新聞　約300万語（β 単位）　昭和41年朝日・毎日・読売三紙の三分の一

(イ)漱石・鷗外など文学作品　約89万語

硝子戸の中（35,000 β），坊っちゃん（53,000 s），行人（150,000 β），三四郎（80,000 β），草枕*（58,000 s）

寒山拾得（4,000 s），高瀬舟（2,500 β），山椒大夫（16,000 s），

雁（45,000 s），青年（50,000 β），渋江抽斎*（150,000 s）

城の崎にて（700 β），焚火（2,400 β）

羅生門（4,000 β），鼻（4,000 β）

遊子方言（7,600 β），浮世風呂（78,000 β），浮世床（50,000 β），心中天

網島（10,000 β），今昔物語集（45,000 β），当世書生気質（50,000 β）

(ウ) 高校教科書* 約60万語 M単位

政治経済，倫理社会，日本史，世界史，地理B，生物I，化学I，物理I，地学I

(エ) 分類語彙表 3.5万語

* 印のものは，現在処理中である。数字の後につけた，β・sは言語単位である。β単位の説明は国研報告12（「現代雑誌九十種の用語用字」）を，s単位の説明は靍岡昭夫「国語研究のための索引作成システム」（「電子計算機による国語研究Ⅷ」）を参照のこと。M単位は高校教科書調査に採用されている言語単位であり，漢語以外は最小単位を一単位とする（漢語は最小単位の一次結合）とする単位である。詳しい説明は後に出る語彙表の説明にゆずる。

2—2 言語処理プログラムの利用

1966年以来われわれは各種の言語処理プログラムを作ってきた。その多くは実験プログラムであったり，使い捨てのプログラムであったりした。この際，これらを一つのシステムの中に組み入れていろいろな言語処理を可能にしよう。以下にあげるものは現在国語研究所が有する各種の言語処理プログラムである。（ ）内は作成者。

各種言語処理プログラム

(ア)	自動単位切り	2種	（石綿・斎藤・木村，江川）
(イ)	よみがな付け	2	（田中，石綿）
(ウ)	かな漢字変換	1	（田中）
(エ)	品詞認定	1	（中野）
(オ)	活用形変換	2	（江川，靍岡）
(カ)	構文解析	4	（石綿・斎藤・木村，中野，佐竹，石綿）
(キ)	KWIC	7	（石綿，斎藤，土屋，斎藤・林，中野，田中卓）
(ク)	WORD COUNT	3	（「新聞」，中野，「教科書」）
(ケ)	漢字調査	1	（野村）

3. 一貫処理システムの説明

一貫処理システムは，国語研究所部内資料LDP—月報別冊4に，第一資料研

究室「語彙調査データの一貫処理法の研究」として，その構想が発表され，重要なサブ・システムである自動単位切り，漢字解読・品詞認定などのプログラムの説明がなされた。本報告はその思想（機械処理と人間作業の複雑なからみあいによる作業工程を，人間による作業をできるだけ機械化して，人間の作業を検査におく）を受け継ぐものであるが，細部にいたっては異なる点も多い。以下，システムの流れ（図1参照）とプログラムの内容について，今回新しくなった点を中心に述べる。

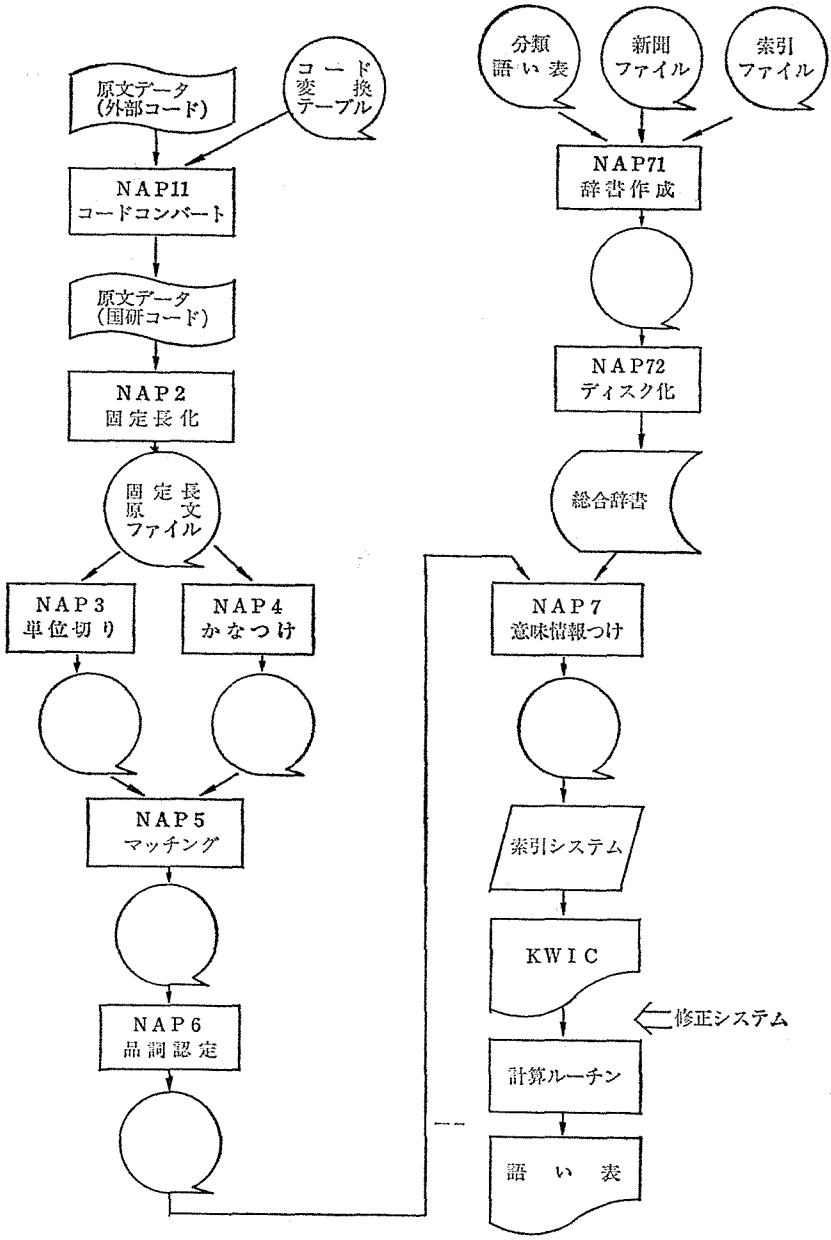
3—1　作業の流れ

図1に示すとおりである。システムを KWIC の作成に重点をおき，その後に処理エラーを人手によって修正し，語彙調査ルーチンに流す。KWIC が出来ていれば，エラーの発見や情報の付加も容易だからである。前のシステムを作ったときには考えられなかった高速漢字プリンタの実現も，KWIC 作成をシステムの中心においた大きな理由の一つである。

総合辞書を利用した各種の情報つけ（本報告では「意味情報」に限ったが，いろいろな情報つけ——たとえば，単位切り，よみがなつけ，品詞情報つけなどにも利用できる）のルーチンを作ったのも今回の新しい試みである。前のシステムでは辞書はできるだけ小さくし，処理はプログラムによっておこなうことを基本においた。これは，処理のスピードをあげること，どんなデータがきても処理できるようにすることのためであった。しかし，現在では高速のディスク装置が利用できること，前述したような大量のデータが利用できるようになったことなどがこのルーチンをもうけた理由である。

単位切りとかなつけのルーチンを並行処理にしたのも新しい点である。処理を直列に並べると処理の誤りが累加的に増えるためである。そういう点では，品詞認定も一緒にすべきかもしれない。というのは，品詞認定と単位切りには次のように処理上の共通点がある。すなわち，字種の違いの利用，テーブルの利用（助詞・助動詞，副詞・連体詞・形式名詞などのテーブルを利用して，単位の認定，品詞情報の付加をおこなう）の二点である。処理の順序が，単位切りは文頭から，品詞認定は文末からおこなう点が異なるが，単位切りを文頭からやらねばならない処理上の理由はないように思われる。したがって，この品

図1　一貫処理（NAP）システム　ブロックチャート



詞認定と単位切りは一つのプログラムにまとめることができる。そうしたほうが処理のスピードや精度をあげることができそうである。しかし，また別々のプログラムにしておくことによって，単位切りされているだけで品詞情報がついていないデータ（この種のデータは相当な量に達する）に品詞情報をつけることができる。今回報告するのは，単位切りとよみがなつけは並行処理，品詞認定はその後において直列処理としたシステムについてである。

国語研究所外のデータを国語研究所コードに直し，利用するルーチンを入れたのも新しい試みである。前のシステムを作った頃は，まだ電子計算機による言語処理が一般には本格的に始まっていなかったのである。

次に各ルーチン・サブシステムについてその処理の内容について述べる。（ ）内はプログラム名称である。

3—2　外部データを国語研究所コードに変換する（NAP 1）

このルーチンは二つに分れる。国語研究所コードと外部データコードとの変換テーブルを作成するプログラム（NAP 10）と，その変換テーブルを用いて，国語研究所外のデータを国語研究所コードに変換するコードコンバートのプログラム（NAP 11）の二つである。

NAP 10用のデータは，現在，写研コード・JICST コード・日電コードと国研（国語研究所の略。以下同。）コードの4種類である。データは図2のようなフォーマットで磁気ディスクに蓄えられる。国研コードには，外部理論コードに対応する国研用文字が入っている。国研コード1は国研コード自身，2は写研コード，3は日電コード，4は JICST コードに対応する文字が入っている。たとえば，外部理論コード16進表示 1234 が，国研コードでは「見」，写研コードでは「省」，日電コードでは「の」，JICST コードでは「横」だとすると，テーブルは「1234見省の横」というようになる。処理は，データを3等分して（理論上のデータ数は4バイト69904種であり，この3等分は 23301 種），それぞれメインメモリー内に展開してコード変換する。したがって，一つのデータを全て変換するには，三度メモリー内での辞書ひきがおこなわれる。

NAP 11は外部コードを国研コードに変換する。NAP 10でのべたように変換は三度おこなわれて完全になる。まず，最初に外部コードが 0000～4FFF の

図2　コード変換テーブル

外部理論コード	国研コード1	国研コード2	国研コード3	国研コード4	
4バイト	2バイト	2バイト	2バイト	2バイト	

データが国研コードに変換され，次に5000～9FFF，最後にA000～FFFF のデータが国研コードに変換される。変換テーブルは3等分されたそれぞれが，外部理論コード自身×2を自分の番地として展開される。たとえば，1234という外部理論コードに対応する文字が「見」だとすると，1234×2＝2468，2469番地に「見」という文字を入れる。また，6789という外部理論コードに対応する文字が「農」だとすると，6789－5000＝1789，1789×2＝3578，3579番地に「農」という文字を入れるという具合である。

3—3　固定長化（NAP2）

入力データ（国研コード，あるいは国研コードに変換された外部データ）は可変長レコードと考えてよい。以降のプログラムで処理しやすくするために，このプログラムではデータを固定長レコードに直す。

3—4　単位切り（NAP3）

江川清「漢字かな混り文の『自動単位分割』に関する一研究」（計量国語学43/44号，1968），同「単位分割自動化のシステムについて」（計量国語学51号，1969）の方法にほぼ従っている。今回の実験では長い単位に切ることだけを目的とする。詳しくは上記論文を参照していただきたい。細かい点で江川方式と異なる。その主なものは，江川は「ら線状」の処理（プログラム内で何回か処理を繰り返して精度をあげる）を行なったが，今回は直線的な処理（一回きりの処理）ですます。いくつかの辞書を利用するが，ここでは辞書の中で優先順位を設け精度を高めている。検索方式はISAM（インデックス・シーケンシャル方式）にならっている。エラー処理したものについてはフィードバックによって修正することができる。以下に処理の概要を簡単に記す。

(1)　字種の判別をおこなう。

(2)　次のものは一字を一単位とし，確定する。

記号類（．，「 」（ ）…）

「を」

(3) 英字・数字・カタカナの処理

・英字連続を一単位とする。ただし，直前・直後が数字のときは，それも加える。

・数字連続を一単位とする。ただし，直後が助数詞（テーブルに定める。一字とする）の場合はこれをつなげる。

・カタカナ連続は一単位とする。

(4) 漢字の処理

・漢字の前で切る。

・他の規則が適用されて，分割されそうな送りがなはテーブルをもうけて処理する。

(5) ひらがなの処理

・メモリー内に展開されたテーブルによる。

テーブルの構成と検索および適用の方法は以下による。

テーブルはインデックス部と辞書部に分かれる。辞書部はデータが入り，インデックス部はデータをいくつかにまとめたそれぞれの先頭の文字と番地が入る。したがって，辞書部内データはその先頭の文字によってソート（50音順配列）されている。

インデックス部の文字は上昇順にソート（50音順配列）されている。

辞書部内データは，同形は文字列の長いものを先におき，これを優先的に適用する。同一インデックス内に入るデータは，同形間で文字列の長いものを先におくだけで，その他には制約はない。したがって，出現率の高い文字列を先におけば処理のスピードが高められる。また，優先的に適用したい文字列があればこれを先にすれば，その指示どおりに分割される。

例をもって示そう。

〔インデックス部〕

あ　001　か　010　さ　020……

〔辞書部〕

いたし	3	こうした	211
いずれ	3	これら	3

あなた	3	ことば	3
あと	2	こと	2
いう	2	これ	2

メモリーの中では，「いたし」以降は１番地以降に，「こうした」以降は10番地以降に配置される。いま入力データ「こうした」がはいってきた場合，データの先頭文字「こ」によって，インデックス部を調べ，辞書部の10番地以降を調べればよいことがわかる。10番地を調べると最長一致で「こうした　２１１」と一致し，分割指示「２１１」を得る。分割指示により，「こう した」は２字・１字・１字に分割すればよいことがわかり，「こう―し―た」と分割される。

辞書部の先においた方を優先するということは，たとえば，入力データ「これら」が，辞書部「これら　3」を先におくことによって，「これ　2」の適用を防ぐという意味をもつ。

単位切りの辞書は，このように単に単語集を辞書にすることだけでは誤った分割されるおそれがでてくるので，文字連続の調査結果を用いて構成することが望ましい。文字連続の調査については，斎藤秀紀「漢字仮名混り文のエントロピー」（計量国語学43／44号 1968）と同「漢字かなまじり文の文字連糸表」(LDP月報別冊８　1971）があり，有用である。

辞書に入れる文字連続は前述単位切り手順により分割されなかった文字連続を正しく分割するために入れるものと，前述の手順によって誤って切られるおそれのあるものをこれで正しく切るものとがある。たとえば，「確かに」や「正しい」は，「に」や「し」によって誤って切られるおそれがあるが，この項目を入れておくことによってその誤りを防ぐという具合である。

3―5　よみがなつけ（NAP４）

田中章夫「漢字かなまじり文を全文カナ書きローマ字書きに変換するシステムについて」（電子計算機による国語研究Ⅱ）のプログラムを使用した。今回，このシステムにのせるためにかえた点は，処理速度をあげるために漢字テーブルをメインメモリー内に展開したことである。

方法の詳しい点は文献にゆずるとして，簡単に方法を説明しよう。

国語研究所の漢字テレタイプ盤内字約2100字について，そのよみがなについ

ての情報をもったテーブルを用意する。このテーブルは三種類にわかれる。

転写回路用テーブル……漢字テーブルのよみがなをそのまま転写するもの。約700字である。

環境演算回路用テーブル……前後の文字の種類により，環境演算をおこない，よみがなを採用する。約500字である。

環境演算は漢字の前後が漢字かかなかによって，論理演算をおこない，その結果によってどのよみがなを取るかを決定する。表1は環境演算の結果を示す。

表1　環境演算の結果

漢字の現われ方	漢字Pの環境		環境演算の結果							
	P−1	P+1	Aa	Bb	Cc	Dd	Ee	Ff	Gg	Hh
前後トモナシ	0	0	01	01	01	01	01	01	01	01
後ダケアリ	0	1	10	01	10	01	10	01	10	01
前ダケアリ	1	0	10	10	01	01	10	10	01	00
前後トモアリ	1	1	10	10	10	10	01	01	01	01

漢字：1　非漢字：0　ヨミガナ記入：0　ヨミガナ無記入：1

テーブルの例

互（aご）（Aたがい）

崩（cほう）（Cくず）

尋（cじん）（Aたず）（Fひろ）

処理の手順を説明しよう。入力文「お互に」の場合，漢字「互」の環境は前後とも漢字なしなので，漢字テーブルの（aご）と（Aたがい）のうち，環境演算結果の指示によりAをとり「お互〔たがい〕に」とよみがなをつける。同様に，入力文「土砂の崩壊，山崩れ」の場合，漢字「崩」の環境は，前者は前が漢字なし後が漢字ありなので，テーブルのうち環境演算結果の指示により，cをとり，後者は前が漢字あり後が漢字なしなので，Cをとる。その結果，出力文は「土砂の崩〔ほう〕壊，山崩〔くず〕れ」となる。

指定環境処理回路用テーブル……前後に特定の文字があらわれた場合だけ特定のよみをとり，その他は環境演算回路と同じ処理をする。約900字。

テーブル例

荷（１ｂか）（２Ｂに）＊M重２／M初３／N担１／N重１

騒（１ｃそう）（２Ｃさわ）＊M物１／Nぎ２

処理の手順を説明しよう。テーブルの＊以降にある漢字が前（Mの場合）または後（Nの場合）にきたとき，指定（数字であらわされている）の読みを＊の前にさがす。入力文「この荷物を運ぶのは重荷だ」という場合，前者の「荷」は後に「物」があり，この漢字が＊以降にないから，環境演算回路によってＢにをとる。後者の「荷」は前に「重」で＊以降にM重２があり，よみ２にをとる。その結果，出力文「この荷〔に〕物を運ぶのは重荷〔に〕だ」をえる。

3—6　マッチング（NAP５）

この処理は，単位切り（NAP３）の出力とかなつけ（NAP４）の出力をあわせ，一つの語によみがながついているという形にするものである。

3—7　品詞認定（NAP６）

筆者「品詞認定の自動化」（電子計算機による国語研究Ⅲ，1971）の方法による。論文では，三つの方法，すなわち辞書による方法，語形による方法，語の接続による方法について述べている。ここでは，辞書による方法は，NAP７の意味情報つけにその可能性を残し，語形による方法と語の接続による方法を採った。したがって，プログラムは二つに分れる。語形による方法（NAP 61）と語の接続による方法（NAP 62）である。

詳しい説明は文献にゆずるが，簡単に処理の概要を記す。

語形による方法では，まず字種の判別をおこなう。次に，助詞・助動詞のテーブル（121語），特殊語のテーブル（漢字書き３語，漢字かなまじり10語，ひらがな書き91語）を調べ，語形が合えばテーブルにある情報を転写する。最後に，語末の文字（１～２字）の判定により，仮の情報をつける。

語末の文字を調べる

１．語末は漢字，カタカナ，英文字，数字→名詞

２．語末は記号→記号

３．語末は「い」→形容詞・終止連体形，動詞・未然連用形

4．語末は「く」→形容詞・連用形，動詞・終止連体形

5．語末は「で」→形容動詞・連用形

6．語末は「に」→形容動詞・連用形

7．語末は「だ」→形容動詞・終止形

8．語末は「な」→形容動詞・連体形

9．語末は「る」→動詞・終止連体形

10．語末は「れ」→動詞・仮定形

11．語末は「よ」→動詞・命令形

12．語末は「かろ」→形容詞・未然形

13．語末は「だろ」→形容動詞・未然形

14．語末は「ろ」→動詞・命令形

15．語末は「かっ」→形容詞・連用形

16．語末は「だっ」→形容動詞・連用形

17．語末は「っ」→動詞・連用形

18．語末は「なら」→形容動詞・仮定形

19．語末は漢字＋ひらがな→動詞

20．語末はイ段→動詞・未然連用形

21．語末はエ段→動詞・未然連用仮定形

22．語末はウ段→動詞・終止連体形

23．語末はア段→動詞・未然形

この方法によると，入力文「広い 門 の 下 で 雨やみ を 待っ て ゐ た 。 」は，「広い（形容詞・終止連体形，動詞・未然連用形） 門（名詞） の（助詞） 下（名詞） で（助詞） 雨やみ（動詞・未然連用形） を（助詞） 待っ（動詞・連用形） て（助詞） ゐ（動詞・未然連用形）た（助動詞） 。（記号）」と品詞認定される。

接続による方法では，語形による方法でつけられた品詞情報を修正する。処理の基本的な考え方は，文中においてある語，とくに助詞，助動詞との語の連続は自由ではなく，かなりの制約があるのは知られているとおりである。その制約をテーブルにして，これにより品詞を決定する。テーブルは次のとおり。

テーブルフォーマット

テーブル1

見出し語	@	情報	@	制限情報(1)							@	制限情報(2)	@	E／i
				#	助詞	……	#	品詞活用情報	／	強制入力情報				

テーブル2

品詞	@	制限情報(1)						@	制限情報(2)	@	E／i
		#	助詞	……	#	品詞活用情報	／	強制入力情報			

テーブル例

テーブル1

の@格助@#と#から#で#へ#より#まで#だけ#ばかり#こそ#など#ぐらい#　　I　+0　／0　@　@Ei

を@格助@#と#から#まで#の#だけ#ばかり#こそ#さえ#すら#のみ#など#ぐらい#　0　／0　@　@E／i

た@助動・過去・た・終止連体@H9　9／H9@　@E／i

テーブル2

X　@#か#さ#ぞ#ね#よ#　　H　+@　@E／i

テーブルフォーマット中，制限情報(1)は見出し語の直前が何であるかを示し，制限情報(2)は見出し語の直後が何であるかを示す。ただし，制限情報(2)は今回は用いない。テーブル例中，X，I，H，9，+，0……などコード化された品詞および活用情報である。これについては文献を参照していただきたい。

例をもって，処理の手順を説明しよう。入力文は語形による方法で品詞認定された文を用いる。入力文（雨やみ（動詞・未然連用形）　を（助詞）　待っ（動詞・連用形）　て（助詞）　ゐ（動詞・未然連用形）　た（助動詞）　・（記号）」は，次のように処理される。

文末の「。（記号）」を取り出し，テーブル2の記号（X）を調べる。処理文記号の直前は「か・さ・ぞ……」の助詞ではなく，活用情報H（終止形）でもない（「た」は終止形なのだが，語形による認定ではそこまで情報がついていない。テーブルには強制入力情報はなく，次の語（直前の語）の処理にうつ

る。「た（助動詞）」をテーブルの中に探し，その中の情報（助詞・過去・た・終止連体）を出力する。その制限情報が処理文直前の語と一致するか調べる。「ゐ（動詞・未然連用形）」とH9（動詞・連用形）と一致しない。そこで，強制入力情報（／の後），H9を強制的につける。以下同様に処理すれば，次の出力を得る。

「雨やみ（名詞）　を（格助詞）　待っ（動詞・連用形）　て（接続助詞）　ゐ（動詞・連用形）　た（助動詞・過去・た・終止連体形）　。（記号）」

3—8　総合辞書作成ルーチン

総合辞書は一貫処理システムの意味情報つけやその他の文法情報つけにも用いられるが，その他の語彙研究・意味研究・文法研究等々いろいろな研究に用いられるように設計された，その名のとおりの総合辞書である。

現在，その内容は「分類語彙表」を中心に，新聞語彙調査の結果や漱石・鷗外の用語索引の見出し語などを含んでいる。将来は，高校教科書の結果の他，一般に使われている国語辞典や英和・和英辞典などや，古典の索引などもそっくり総合辞書の中にとりこみ，広範囲の利用に供したい。

データのフォーマットは次のとおりである。

通し番号	漢字かなまじり見出し	同一見出しの番号	かな見出し	文法情報	意味情報	出現率	出典	配列情報

現在の収録語数は約7万である。

総合辞書は磁気ディスクに蓄えられ，検索方式は ISAM である。

3—9　意味情報つけ（NAP7）

総合辞書によって意味情報（分類語彙表の番号）をつけるプログラムである。また，品詞認定やかなつけのプログラムでつけられなかった品詞情報やかな情報，活用形変換のための活用情報もここでつける。

3—10　索引システムやワードカウント・ルーチンへの転換

以上で，一貫処理システムの主な処理が終った。この後は，「索引作成のためのプログラムライブラリ」によって，KWIC や語彙表をつくる。

ただし，後に示すように，このシステムでは残念ながら100%の正解率は得られていない。処理を誤った部分については，KWIC を見ることによってそ

のエラーの部分を発見し，人手で修正する。また，エラーがなくても語彙調査ルーチンにわたすために同語異語の判別情報を付加しなければならない。

これらの修正ルーチンや同語異語の判別情報付加のルーチンは「索引作成のためのプログラムライブラリ」のルーチンを使えばよい。しかし，漢字ディスプレイによる修正など，なお修正方式の効率化をはかる必要がある。

3—11　テストランの結果

「電子計算機による国語研究Ⅶ」は電算写植によって印刷された。そこで印刷会社に頼み，その中間出力である紙テープを手に入れた。この紙テープをテストデータにした。すなわち，電算写植用の写研コードで打たれた紙テープがこのテストランの入力データである。

処理結果を図2～7に示す。

処理の精度と処理エラーの原因は次のとおりである。

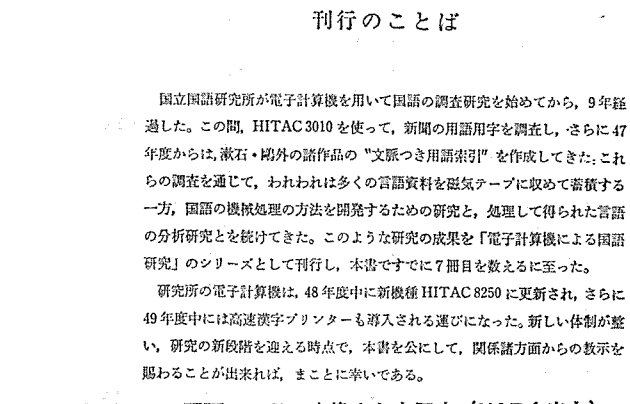
(1)　コード変換　　コード変換自体のエラーは無い。しかし，電算写植用につけられたポイント情報・ページ情報・改行情報・ルビ情報等の無視によって処理エラーが起こることはありえる。また，電算写植においては最終結果は印刷物である。したがってその中間結果である紙テープにパンチエラーがあっても最終的に印刷物が正しければ（切り貼りをすることによって，パンチエラーを修正するなど）よい。このようなエラーが一ケ所（「漱」が「瀬」になっている）あった。また，人の目で見て正しければよいものが四ケ所（漢数字の「一」であるべきところ，カタカナ長音であるべきところをそれぞれマイナスで代用した……「一方」「テープ」「シリーズ」「プリンター」）あった。また，写研コードの盤内字が国研コードの盤外字であるもの，写研コードの盤外字が国研コードの盤内または盤外字であるものの処理をしていないための処理エラーが一ケ所（「鷗」）あった。これは，コードコンバートでは常に考えなければならない重要な問題であるが，テストランでは放置した。

(2)　単位切り　　142語に切れるべきところ，32ケ所に処理エラーがあった。その原因は次のとおりである。

数字が関係するところ……6ケ所

「9　年経過し」「47　年度」のように，助数詞のテーブルを設ければ正し

図2　入力原文　電算写植による印刷例
（「電子計算機による国語研究Ⅶ」）



刊行のことば

国立国語研究所が電子計算機を用いて国語の調査研究を始めてから，9年経過した。この間，HITAC 3010 を使って，新聞の用語用字を調査し，さらに47年度からは，漱石・鴎外の諸作品の "文脈つき用語索引" を作成してきた。これらの調査を通じて，われわれは多くの言語資料を磁気テープに収めて蓄積する一方，国語の機械処理の方法を開発するための研究と，処理して得られた言語の分析研究とを続けてきた。このような研究の成果を「電子計算機による国語研究」のシリーズとして刊行し，本書ですでに7冊目を数えるに至った。

研究所の電子計算機は，48年度中に新機種 HITAC 8250 に更新され，さらに49年度中には高速漢字プリンターも導入される運びになった。新しい体制が整い，研究の新段階を迎える時点で，本書を公にして，関係諸方面からの教示を賜わることが出来れば，まことに幸いである。

図3　入力データ・国研コードに変換された原文（NAP 1 出力）

国立国語研究所が電子計算機を用いて国語の調査研究を始めてから９年経過した．この間，ＨＩＴＡＣ３０１０を使って，新聞の用語用字を調査し，さらに４７年度からは瀬石・●外の諸作品"文脈つき用語索引"を作成してきた．これらの調査を通じてわれわれは多くの言語資料を磁気テープに収めて蓄積する一方国語の機械処理の方法を開発するための研究と処理して得られた言語の分析研究とを続けてきた．このような研究の成果を「電子計算機による国語研究」のシリーズとして刊行し本書ですでに７冊目を数えるに至った．研究所の電子計算機は４８年度中に新機種ＨＩＴＡＣ８２５０に更新されさらに４９年度中には高速漢字プリンターも導入される運びになった．新しい体制が整い研究の新段

図4　自動単位切り済データ（NAP 3 出力）

国立国語研究所 が 電子計算機 を 用い て 国語 の 調査研究 を 始め て から ９ 年経過し た ． この 間 ， ＨＩＴＡＣ ３０１０ を 使っ て ， 新聞 の 用語用字 を 調査し ， さら に ４７ 年度 から は 瀬石 ・ ●外 の 諸作品 " " 文脈つ き 用語索引 " " を 作成 し て きた ． これ ら の 調査 を 通じ て われわれ は 多く の 言語資料 を 磁気テ ー プ に 収め て 蓄積 する 一 方国語 の 機械処理 の 方法 を 開発 する ため の 研究 と 処理 し て 得られ た 言語 の 分析研究 と を 続け て きた ． この ような 研究 の 成果 を 「 電子計算機 に よ る 国語研究 」 の シリ ー ズ と し て 刊行し 本書 です で に ７ 冊目 を 数える に 至っ た ． 研究所 の 電子計算機 は ４８ 年度中 に 新機種 ＨＩＴＡＣ ８２５０ に 更新されさら に

図5　自動かなつけ済データ（NAP 4 出力）

国【こく】立【りつ】国【こく】語【ご】研【けん】究【きゅう】所【しょ】が電【でん】子【し】計【けい】算【さん】機【き】を用【もち】いて国【こく】語【ご】の調【ちょう】査【さ】研【けん】究【きゅう】を始【はじ】めてから9年【ねん】経【けい】過【す】した．この間【あいだ】，ＨＩＴＡＣ３０１０を使【つか】って，新【しん】聞【ぶん】の用【よう】語【ご】用【よう】字【じ】を調【ちょう】査【さ】し，さらに47年【ねん】度【ど】からは瀬【せ】石【せき】・●【NONE】外【がい】の諸【しょ】作【さく】品【ひん】"文【ぶん】脈【みゃく】つき用【よう】語【ご】索【さく】引【いん】"を作【さく】成【せい】してきた．これらの調【ちょう】査【さ】を通【とお】じてわれわれは多【おお】くの言【げん】語【ご】資【し】料【りょう】を磁【じ】気【き】テープに収【おさ】めて蓄【ちく】積【せき】する一方【ほう】国【こく】語【ご】の機【き】械【かい】処【しょ】理【り】の方【ほう】法【ほう】を開【かい】発【はつ】するための研【けん】究【きゅう】と処【しょ】理【り】して得【え】られた言【げん】語【ご】の分【ぶん】析【せき】研【けん

図6　自動品詞認定・意味情報つけ済データ（NAP７出力）

```
000001 00001 001 国立国語研究所    こくりつこくごけんきー1ーーーー
000002 00001 002 が                が                -R13--4・113
000003 00001 003 電子計算機        でんしけいさんき  -1-----
000004 00001 004 を                を                -R1---
000005 00001 005 用い              もちい            -EG29-
000006 00001 006 て                て                -R2---
000007 00001 007 国語              こくご            -1----1・3101
000008 00001 008 の                の                -R1---1・100
000009 00001 009 調査研究          ちょうさけんきゅう -1-----
000010 00001 010 を                を                -R1---
000011 00001 011 始め              はじめ            -Z-----
000012 00001 012 て                て                -1--%-
000013 00001 013 から              から              -R13--
000014 00001 014 9                 9                 -X-----
000015 00001 015 年経過し          ねんけいすし      -E--9-
000016 00001 016 た                た                -P-1+-
000017 00002 017 ．                ．                -Y-----
000018 00002 001 この              この              -Z----3・100
000019 00002 002 間                あいだ            -1----1・1610
000020 00002 003 ，                ，                -Y-----
```

図7　ＫＷＩＣ（索引システム出力）

```
間              002 02 X -1----- 1U16100 年経過した．この 間 ，HITAC30
多く            003 10 X -1----- 1U19800 通じてわれわれは 多く の言語資料を磁
収め            003 18 X -EID9-  -------0 料を磁気テープに 収め て蓄積する一方
が              001 02 X -R13--  4U113●0   国立国語研究所 が 電子計算機を用い
開発            003 29 X -1----- 1U38220 機械処理の方法を 開発 するための研究
から            001 13 X -R13--  -------0 調査研究を始めて から 9年経過した．
から            002 20 X -R13--  -------0 ，さらに47年度 から は瀬石・●外の
き              002 36 X -EK39-  -------0 索引"を作成して き た．これらの調査
き              002 29 X -ZK3--  -------0 の諸作品"文脈つ き 用語索引"を作成
機械処理        003 25 X -1----- -------0 積する一方国語の 機械処理 の方法を開
研究            003 33 X -1----- 1U30650 を開発するための 研究 と処理ギ
言語資料        003 12 X -1----- -------0 われわれは多くの 言語資料 を磁気テー
国語            001 07 X -1----- 1U31010 子計算機を用いて 国語 の調査研究を始
国立国語研究所  001 01 X -1----- -------0                国立国語研究所 が電
この            002 01 X -Z----- 3U100●0 ら9年経過した．この 間，HITAC
こい            003 01 X -E--Q-  1U100●0 を作成してきた．これ らの調査を通じ
作成            002 33 X -1----- -------0 つき用語索引"を 作成 してきた．これ
さら            002 16 X -1----- 3U16610 語用字を調査し，さら に47年度から
```

く処理できるもの，「HITAC 8250」などのように英数字連続を切り離すための処理エラーなどが含まれる。

長音番号が関係するもの……3ケ所

これは，原データパンチミスによる。

漢字連続であるべきところ……2ケ所

原データのパンチミスと，副詞の漢字書きと名詞の漢字書きの連続のために切り離せなかったもの（「一　方国語　の」）である。後者は，テーブルに副詞の漢字書きリストを入れれば正しく処理できる。

テーブルに原因があるもの……6ケ所

「蓄積　する」「さら　に」「本書　です　で　に」と切り離されたのは「する」「に」「です」「で」がテーブルにあったためである。「さらに，ですでに」をテーブルに入れれば解決する。「する」は，これをテーブルからはずすと，「びっくりする」などが一語になってしまい，どうするか難しい問題である。

(3)　よみがなつけ　　84字中4字のよみまちがいである。「経〔けい〕過〔す〕した」「調査を通〔とお〕じて」「一方〔ほう〕」のうち，前の二例は漢字テーブルの修正で解決する。後者は連濁の問題でありテーブルを直しはじめるとテーブルの量が大きくなりすぎ解決が難しい。

(4)　品詞認定　　48語中9語のつけまちがいである。このうち5語は単位切りエラーによる。「この（不明）間」「調査し（不明），」「始め　て（名詞）から」はテーブルを直すことによって正しくなる。

(5)　意味情報つけ　　異なりで37語中21語に情報がつかなかった。

このうち，総合辞書が短い単位で登録されているのに，本実験データは長い単位の処理であるための未処理が7例，正しく活用変換がされていないものが4例，助詞・助動詞・記号につかないものが5例，前処理のエラーによるものが2例であった。「ため・作成」に情報がつかなかったのは総合辞書のエラーによる。

以上の結果，現在では約8割が正しく処理されており，なおシステムの能力アップをはかることと修正方法の簡単化をはかることによって充分実用に供す

ることができると思われる。

4. おわりに

今回の実験は一貫処理システムの開発のための第一段階と考えている。今後，辞書を多用して処理する方式を使うことや，構文解析プログラムの利用によってその精度をあげたい。また，同語異語判別のアルゴリズムを研究し，その自動化についても研究する必要がある。

一貫処理のシステムは，もともと実用化をねらって研究・開発が進められた。その点においては，精度が90%近くになれば当初の目的が達せられ，修正システムを導入して十分採算がとれるものと考える。とくに少量の調査についても計算機の利用が可能になるだろう。

一貫処理を可能ならしめるためには，各種の言語処理プログラムと多量の言語データ，および人的資源が用意されていなければならない。そのためには，各機関との協力態勢を作ること，特に国語研究所の果さねばならない役割は大きいと思われる。

言語処理の発展過程を次の三時期に分けると，このシステムは第二期のものであると考えられる。

第一期　多くの人的作業を加えて計算機処理を可能にする時代。

第二期　言語的な作業の多くを計算機に肩がわりさせ，人間でしかできない面を人間が行なう。機械と人間の調和の時代。

第三期　完全自動処理の時代。

完全な自動処理を実現するには，なお各種の言語研究や分析手法，処理法の開発が行なわれなければならない。そのような分析・研究にも一貫処理システムが利用できるものと信じる。

このシステムは大阪外国語大学教授田中章夫氏（元国立国語研究所所員）・国立国語研究所員江川清氏等々の多くの人々の研究の上に完成したものである。

また，実験にあたって，ファコムハイタック株式会社今井良一・中島保行両

氏の協力があった。プログラムの作成・データの整理等には研究補助員長田厚子嬢の協力がなければ，このシステムの完成はまだまだ見られなかっただろう。

以上，多くの人々に感謝の意を表するものである。

この報告は，昭和50年度国立国語研究所研究発表会「用語用字調査と機械処理」（昭和51年3月24日岩波ホール）において発表したものにもとづいている。

参考文献

〔国研内言語処理文献〕

1．第一資料研究室「語彙調査データの一貫処理法の研究」（LDP 4，1969）
2．石綿敏雄・斎藤秀紀・木村繁「言語単位分割自動化の研究」（計量国語学 50，1969）
3．江川清「漢字かな混り文の『自動単位分割』に関する一研究」（計量国語学 43／44，1968）
4．——「単位分割自動化のシステムについて」（計量国語学 51，1669）
5．田中章夫「漢字かなまじり文を全文カナ書き・ローマ字書きに変換するシステムについて」（電子計算機による国語研究Ⅱ，1969）
6．——「ヨミガナ方式によるカナ（ローマ字）の漢字変換」（計量国語学 55，1970）
7．中野洋「品詞認定の自動化」（電子計算機による国語研究Ⅲ，1971）
8．江川清「『活用形処理』の自動化に関する一方式」（電子計算機による国語研究Ⅱ 1969）
9．靏岡昭夫「文語形・口語形活用語の代表形の変換処理について」（電子計算機による国語研究Ⅴ，1973）
10．石綿敏雄「構文解析自動化の研究Ⅰ」（電子計算機による国語研究Ⅱ，1969）
11．木村繁「構文解析自動化の研究Ⅱ」（電子計算機による国語研究Ⅱ，1969）
12．佐竹秀雄「構文解析の一つの試み」（計量国語学 62，1972）
13．中野洋「構文自動解析の試み」（計量国語学 71，1974）
14．石綿敏雄「変形とその逆探知を含む構文解析」（電子計算機による国語研究 Ⅷ，1976）
15．斎藤秀紀「電子計算機と漢テレによる用語総索引の作成」（電子計算機による国語研究，1968）
16．石綿敏雄「新聞用語調査の用例印字プログラム “COBOL-KWIC”」（電子計算機による国語研究Ⅲ，1971）
17．土屋信一「カナ入力による日本語文総索引の作成」（電子計算機による国語研究Ⅳ，1972）
18．中野洋「索引作成のためのプログラムライブリ」（電子計算機による国語研究Ⅷ，

1976)
19. 中野・斎藤・米田・白木・竹内「高校教科書用語用字調査システム（中間報告)」(季報1975冬)
20. 石綿敏雄「電子計算機による語彙調査の一実験」(ことばの研究Ⅱ, 1965)
21. 斎藤秀紀「電子計算機による語彙調査, Ⅱ, Ⅲ」(電子計算機による国語研究Ⅱ, Ⅲ, Ⅴ, 1969, 71, 73)
22. 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, 1968)
23. 石綿敏雄「COBOL による漢字索引の作成」(電子計算機による国語研究 Ⅱ, 1969)
24. 野村雅昭「新聞漢字調査の機械処理システム」(電子計算機による国語研究Ⅲ, 1971)
25. 斎藤秀紀「漢字プリンターを使用したターンアラウンドシステム, Ⅱ」(電子計算機による国語研究Ⅵ, Ⅶ, 1974, 1975)
26. ――「言語処理におけるターンアラウンド・システム」(電子計算機による国語研究Ⅷ, 1976)
27. 田中卓史「KWIC ・語彙表システム――カード入力磁気ディスク利用――」(季報1977秋)