言語処理における一貫処理システム

言語処理における一貫処理システム

中野　洋

I

電子計算機を利用する最大の利点は，計算機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。

ところが，言語処理の場合集計処理は計算機向きだが，言語的な情報の付加が複雑で，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── たとえば単位切り，よみがなつけ，語種・品詞・活用情報・意味情報の付加など ── には，時間も費用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採算がとれるが，少量の処理などでは，かえって人手でやった方が能率的だという場合がしばしば起こる。これでは計算機本来の利点が損われていることになる。

そこで，現在行っているような言語的な情報付加の作業をできるだけ少なくして，これを計算機に肩がわりさせる方法を考えた。一貫処理システムとはそのような人的作業軽減のための言語処理システムである。

II

1. 人的作業の軽減

(1) 知的作業のみを人間が行うこと……たとえば，清書やフォーマット変換，簡単な単位切りや情報付けは，これを人間の作業から除いて計算機にやらせてもできることである。人間は人間にしかできない知的作業に専念することによって調査の精度を高めることができる。たとえば，語彙調査における同語異語判別などがこれである。

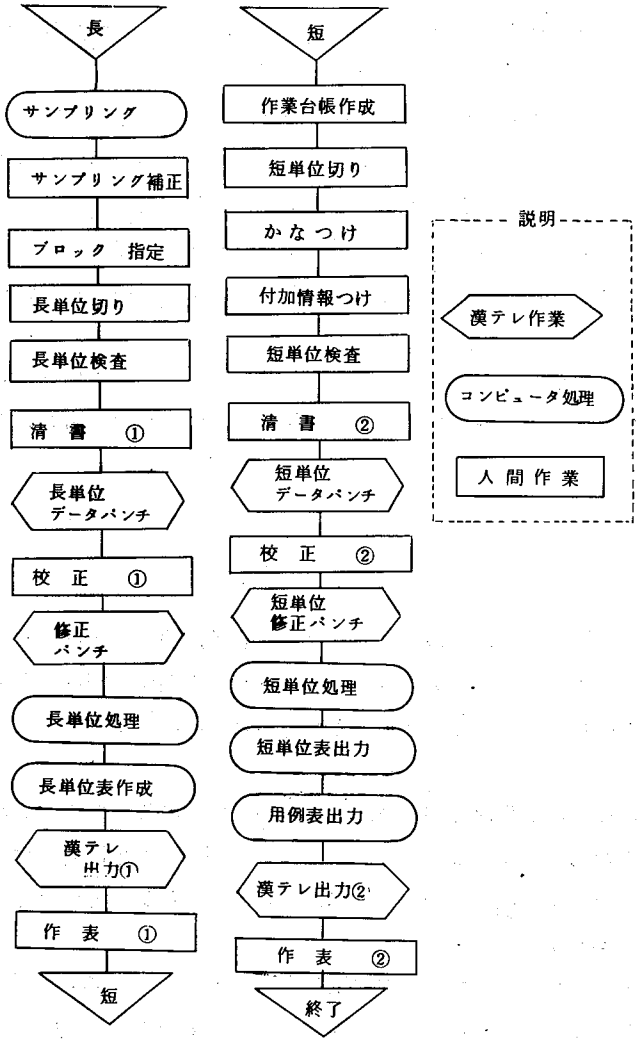
(2) ミスを少なくすること……人間の作業は高度な知的作業が可能だが，同時に簡単なミスもたびたび起こる。すべてを人間の作業によるのならこのような簡単なミスは，人間がまた簡単に発見することができる。しかし，これを機械に任せると，ミスはミスのままで処理されて直されることがなく，処理の精度を落とすばかりか，修正に人的作業をかけることになる。

(3) プレエディットからポストエディットへ……これまでの語彙調査は図1のようにプレエディット方式であり，その作業は膨大なものであった。一貫処理システムではこれを図2のように人的作業を処理の後に持って行き，ここで集中的に作業を行う。こうすることによって作業の受け渡しミスをなくすことができる。

(4) パンチ量を減らすこと……人的作業の結果はすべてパンチされ計算機に入力される。人的作業が少なければ当然パンチ量が減る。これらはすべて機械が行う。

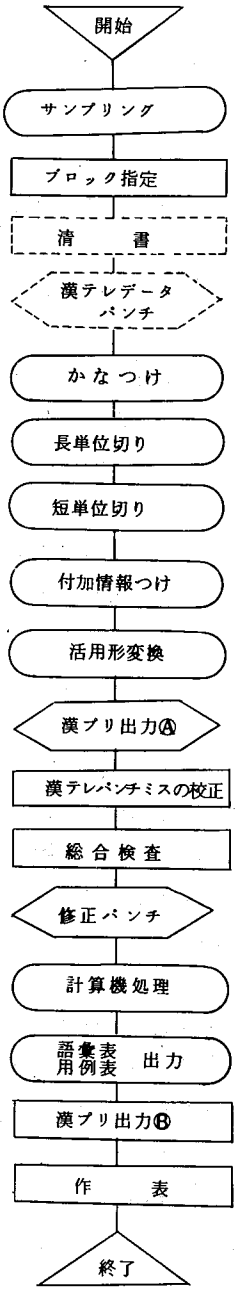
2. 蓄積データ，プログラムの利用

図1　新聞の語彙調査



「語彙調査データの一貫処理法の研究」（LDP4）による。

図2　一貫処理法



(1) 国研外データの利用……電算写植による印刷は，最近急速に増えている。関係者から伺うところによると，1985年には印刷業界の8割は電算写植になるという予想があるそうである。ところで，電算写植の中間出力として言語データがパンチされた紙テープがある。この紙テープを入手し，このシステムに接続することができるなら，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。

(2) 国研内蓄積データの利用……1966年に導入された国研電子計算機の総処理量は後に示すように延べ430万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。

国研データー覧

(ア) 新聞　約300万語　β単位

昭和41年朝日・毎日・読売三紙の文章

(イ) 漱石・鷗外など文学作品　約76万語

漱石　　硝子戸の中（35,000β），坊ちゃん（53,000 S），行人（150,000β），三四郎（80,000β），草枕（58,000 S）

鷗外　　寒山拾得（4,000 S），高瀬舟（2,500β），山椒大夫（16,000 S），雁（45,000 S）

青年（50,000β），*渋江抽斎（150,000 S）

志賀直哉　　城の崎にて（700β），焚火（2,400β）

芥川竜之介　　羅生門（4,000β），鼻（4,000β），その他

その他　　遊子方言（7,600β），浮世風呂（78,000β），浮世床（50,000β），心中天網島（10,000β），今昔物語集（45,000β），当世書生気質（50,000β）

(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，数学Ⅰ

*印のものは，現在処理の途中である。

(3) 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あらゆる言語処理を可能ならしめ，その余った時間を本当の意味での言語処理が現在当面している問題に向けるべきである。

各種言語処理プログラム

(ア)	自動単位切り	2種
(イ)	よみがな付け	2〃
(ウ)	かな漢字変換	1〃
(エ)	品詞認定	1〃
(オ)	活用形変換	2〃
(カ)	構文解析	3〃
(キ)	KWIC	6〃

(ク) 語彙調査　　3種

(ケ) 漢字調査　　1〃

詳しくは参考文献参照

Ⅲ

システムについては，旧第一資料研究室が「語彙調査データの一貫処理法の研究」(LDP4，1969)に発表した。ここでは前システムと異なる点について説明する。

前システムと最も大きく異なる点は，意味情報つけのプログラムを加えたことである。語彙調査システムにおいて，同語異語判別はなくてはならない処理である。しかし，現在のところその自動化は実現していない。このプログラムは辞書の意味情報を付加するだけであるが，将来は自動意味情報つけを試みたい。

つぎに，KWIC出力をメインに置いたことである。これは高速漢字プリンタの利用によって漢字かなまじり文出力が速くなったことと，これが人間による同語異語判別になくてはならないためである。

プログラムはほとんどがCOBOL言語で書かれている。他機種の計算機によっても処理が可能になる。

前システムでは，語種情報・活用情報・語構成情報の付加も自動化することになっているが，今回の実験では，これを入れていない。

また，処理の順序も変わっている。たとえば，単位切りと読みがなつけを並行処理にした。これは精度を落さないためである。

次に，主要プログラムの変更点について述べる。前システムの各プログラムについては参考文献にあたられたい。今回の実験プログラムの内容についての詳細は，別に機会を得て発表したい。

(1) 自動単位切り　　江川方式による。今回の実験では長単位切りのみを扱う。処理法において，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その精度をあげる）を行ったが，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，ここでは辞書の中に優先順位を設け精度を高めている。検索方式はISAM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。

(2) よみがなつけ　　田中方式による。今回の実験では，辞書をメモリー内にもち処理速度をあげている。

(3) 品詞認定　　筆者のプログラム方式による。

(4) 活用形変換　　鶴岡方式による。活用情報は総合辞書から取り出す。

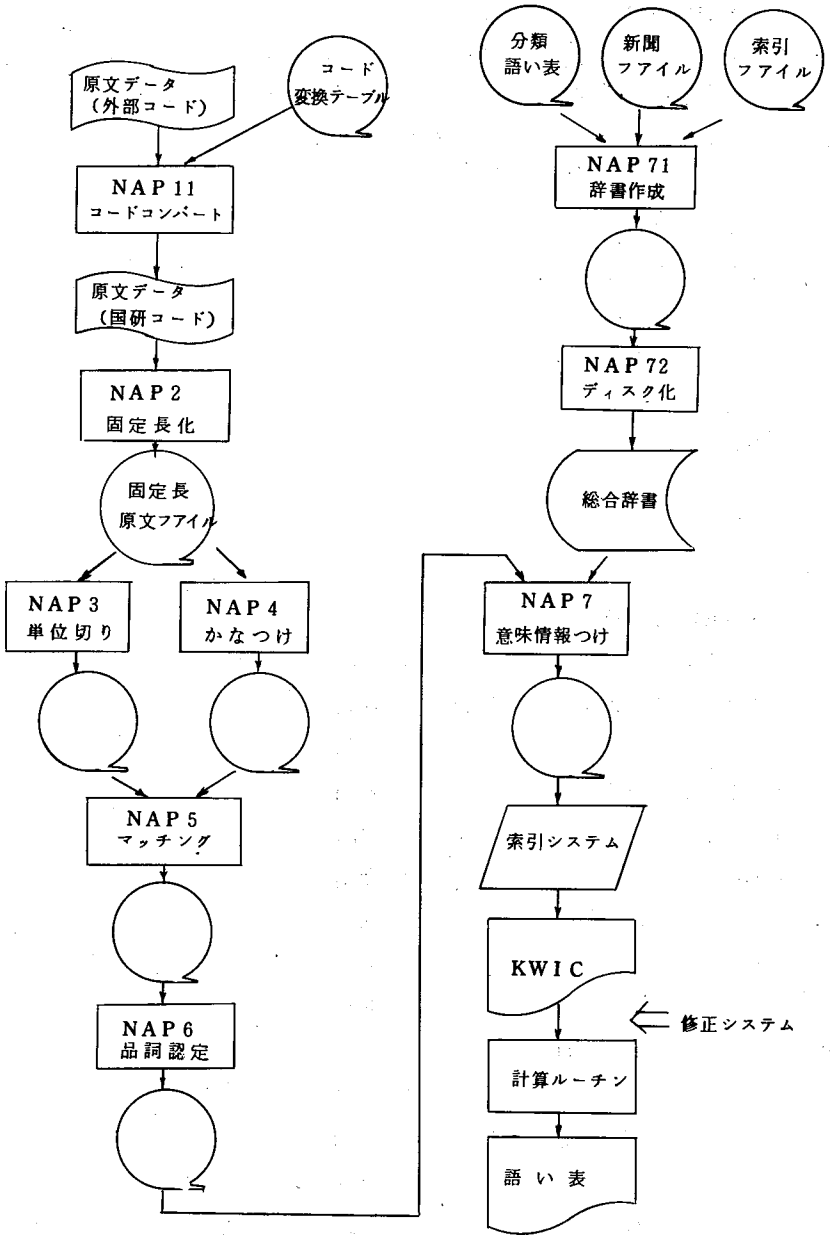
(5) 意味情報つけ　　図3に示したように，『分類語彙表』をもとに，各種のデータを加え，総合辞書を作った。ここでは，ディスクに登録された総合辞書によって，分類番号など各種の情報をつける。

〔処理結果例〕

(原文データ・NAP11またはNAP2入力)

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。

図3　一貫処理（NAP）システム　ブロックチャート



（単位切り済みデータ・NAP3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 。

（かなつけ済みデータ・NAP4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〔か〕した。

（マッチングデータ・NAP5出力）

〔見出し〕	〔よみがな〕
国立国語研究所	こくりつこくごけんきゅうしょ
が	が
電子計算機	でんしけいさんき
を	を
用い	もちい
て	て
国語	こくご
の	の

（品詞認定済みデータ・NAP6出力）

〔見出し〕	〔よみがな〕	〔品詞〕
国立国語研究所	こくりつこくごけんきゅうしょ	名詞
が	が	助詞
電子計算機	でんしけいさんき	名詞
を	を	助詞
用い	もちい	動詞
て	て	助詞
国語	こくご	名詞
の	の	助詞

（意味情報つけ済みデータ・NAP7出力）

〔見出し〕	〔よみがな〕	〔品詞〕	〔分類番号〕
国立国語研究所	こくりつこくごけんきゅうしょ	名詞	
が	が	助詞	
電子計算機	でんしけいさんき	名詞	
を	を	助詞	
用い	もちい	動詞	
て	て	助詞	
国語	こくご	名詞	1.3101

の　　の　　助詞

（KWIC 出力）

〔見出し〕	〔品詞〕	〔分類番号〕	〔文・語番号〕	〔用例〕
研究	名詞	1.3065	003 30	開発するための 研究 と，処理して得
	名詞	1.3065	004 04	た。このような 研究 の成果を「電子
	名詞	1.3065	006 06	い体制が整い， 研究 の新段階を迎え
国語	名詞	1.3101	001 07	計算機を用いて 国語 の調査研究を始
	名詞	1.3101	003 21	蓄積する一方， 国語 の機械処理の方
を	助詞		006 25	方面からの教示 を 賜わることが出来
	助詞		001 10	国語の調査研究 を 始めてから，9年
	助詞		006 09	，研究の新段階 を 迎える時点で，本
	助詞		004 07	な研究の成果 を 「電子計算機によ
	助詞		003 43	の分析研究と を 続けてきた。この

（語い表　出力）

〔順位〕	〔見出し〕	〔品詞〕	〔度数〕	〔全体使用率〕	〔部分使用率〕
1	,	記号	14	70.7 ‰	
2	の	助詞	14	70.7	
3	に	助詞	11	55.6	
4	て	助詞	9	45.5	
5	を	助詞	7	35.4	
50	研究	名詞	3	15.2	29.1
50	電子計算機	名詞	3	15.2	29.1

Ⅳ

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，辞書方式の併用や，構文解析プログラムの利用によって，その精度を上げたい。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。現在，国語研究所はこの条件を満足できる唯一のグループであると思う。

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考えている。

今回の実験にあたって，日立製作所今井良行・中島保行両氏の協力があった。記して感謝の意を表する。

国研内言語処理文献

1. 第一資料研究室（1969）語彙調査データの一貫処理法の研究。（LDP4）
2. 石綿敏雄・斎藤秀紀・木村繁（1969）言語単位分割自動化の研究。（計量国語学 50）
3. 江川清（1969）単位分割自動化のシステムについて。（計量国語学 51）
4. 田中章夫（1969）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）
5. ――（1970）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）
6. 中野洋（1971）品詞認定の自動化。（電子計算機による国語研究Ⅲ）
7. 江川清（1969）「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）
8. 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。（電子計算機による国語研究Ⅴ…以下「国語研究Ⅴ」と表わす）
9. 石綿敏雄（1969）構文解析自動化の研究Ⅰ。（電子計算機による国語研究Ⅱ）
10. 木村繁（1969）構文解析自動化の研究Ⅱ。（電子計算機による国語研究Ⅱ）
11. 佐竹秀雄（1972）構文解析の一つの試み。（計量国語学 62）
12. 中野洋（1974）構文自動解析の試み。（計量国語学 71）
13. 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。（電子計算機による国語研究）
14. 石綿敏雄（1971）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）
15. 土屋信一（1972）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）
16. 中野洋（1975）用語検索システムについて。（季報 1975 秋）
17. 中野・斎藤・米田・白木・竹内（1975）高校教科書用語用字調査システム（中間報告）（季報 1975 冬）
18. 石綿敏雄（1965）電子計算機による語彙調査の一実験。（ことばの研究Ⅱ）
19. 斎藤秀紀（1969，71，73）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）
20. 田中章夫（1968）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）
21. 石綿敏雄（1969）COBOLによる漢字索引の作成。（電子計算機による国語研究Ⅱ）
22. 野村雅昭（1971）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）
23. 斎藤秀紀（1974）漢字プリンターを使用したターンアラウンドシステム。（電子計算機による国語研究Ⅵ）