品詞認定の自動化

著者	中野 洋
雑誌名	電子計算機による国語研究
巻	3
ページ	98-120
発行年	1971-03
シリーズ	国立国語研究所報告 ; 39
URL	http://doi.org/10.15084/00001007

品詞認定の自動化

中野　洋

1. 目的

現在，国立国語研究所の第一資料研究室，第三資料研究室，言語計量調査室では，電子計算機を使った語彙調査を行っている。この調査のシステムは長単位処理と短単位処理の二つに大きく分かれる(注1)。長単位処理の前処理として，長単位切り，短単位処理の前処理として短単位切り，読みがな付け，付加情報付けがある。語彙調査に電子計算機を導入した大きな理由は，短時間で多くのデータを処理できることであった。ところが，この調査を始めてみると，計算機処理に入る前の前処理に多くの時間とエネルギーを費すことがわかった。しかも，エラーの原因の多くがこの人手作業にあった。そこで，前処理も電子計算機にさせるシステムが考えられた。これを我々は，語彙調査一貫システムと呼んでいる(注2)。

語彙調査一貫システムを完成させるためには，自動単位切り(注3)，自動漢字解読(注4)，自動品詞認定のプログラムができていることが必須条件となる。ここでは，その自動品詞認定のプログラムについて説明する。

このプログラムを開発した第一の目的は，一貫処理システムの一部分を受け持つことであった。この方法を考えるにあたって，参考になったのは我々人間の品詞認定における行動であった。品詞認定における我々の行動のうちどれが一番効いているのか，どういう順序で働いているのか，それらをプログラミングすることによってこのような問題を解明する手掛かりを得ることができる。これが，このプログラムを開発する第二の目的であった。その他このプログラムを作っていく過程において得られたものは多い。これらについては順次述べることにしよう。

このプログラムを開発することによって，次のような利点が得られる。

1）語彙調査に計算機を用いたのは短時間に大量のデータを処理できるからであった。ところが手作業がその速度を機械の速度から人間の速度に落してしまった。これを再び，機械の速度に引き上げることができる。

2）品詞認定は誰れにでもできるという作業ではない。何らかの訓練を受けた人が行なうことになる。ところが，ここで二種類の間違いが起こる。一つは，大部分の品詞は単純に付けられるが，そこで簡単な繰り返し作業によるケアレスミスが起こる。他の一つはどの品詞を付ければ良いか解からない場合のミスである。これは専門家に任せなければならない。いずれにせよ，多くの時間を費し，ミスを生むのである。その上，いけないことに人間が付けた場合，どこにミスがあるか解からない。検査を繰り返してもこのようなミスを完全になくすことは難しい。

計算機にやらせた場合，確かに人間が付けたより精度は悪いだろう。しかし，短時間に大量のデータを処理できること，ケアレスミスがないこと，間違いの現われ方が一定すること，つまりどこに間違いがあるかわかることなどの利点がある。

3）前に述べたことだが，このプログラムを使って人間がどのようにして品詞認定をしているのか，どの要素が効いているのか，その順序はどうかなどの実験をすることができる。一種の品詞認定における人間行動のシミュレーションである。

2. 方法の概略

品詞認定の方法はいくつか考えられる。言語情報処理においては何らかの品詞認定を行なっているものが多い。しかしながら，それらは品詞認定だけが目的ではなくて，それを使って，何らかの作業を行なうものである。方法も目的に従って変わる。ここでは国立国語研究所の計算機システムを用いて，語彙調査のための品詞認定を考えることにする。

① 構文解析による方法

構文的関係がその品詞を決定する場合は多い。従って正しい構文の解析による品詞認定はより正しく，かつ詳しくなされる。語彙調査にどの程度の品詞情報が付くかによって変わるが，構文を解析してまでの詳しい情報は，普通，語彙調査に必要ではないようだ。この方法は処理時間が長くなるので大量データの処理には不適当だと思われる。

② 辞書による方法

この方法は普通に用いられている。この方法を実行するには，完全な辞書が必要であること，ランダム・アクセス装置などがあることが望ましい。現在，進行中の語彙調査の結果がそのまま辞書になり心配はない。高速外部記憶装置については，国語研究所の場合，磁気テープ装置しかない。一語一語について辞書を調べるこの方法では，ずいぶんの時間がかかるものと思われる。（出力形式を選べばかなりの速度に上げることができる。）

③ ここで用いる方法

先の二つの方法もある程度含むが，主に我々が品詞認定をどのようにしているかを考え，それをプログラム化したものである。

1）辞書による方法

たとえば，学校で品詞を付けよという宿題がでたとする。その時の最も簡単な方法は，ノートを見るとか，国語辞書を調べるとか，文法書を調べるとかする。時には誰れかに聞く。

この方法は計算機にのせることができる。ノートや国語辞書や文法書，何れかの知識をひとつにまとめ（言葉——品詞）というように整理する。これを磁気テープにして，計算機の辞書とするのである。計算機は，入力された語を一つずつ見，それが磁気テープにあるかを調べる。あれば辞書に載っている品詞をその語に付ければよい。なければ仕方がないから，「不明」の情報でも付けておく。

2）語形による方法

中学校にでも行くと国語の時間に語の活用形について習う。サ行五段動詞「話す」の語尾は「サ，ソ，シ，ス，ス，セ，セ」とか，ア行上一段は「イ，イ，イル，イル，イレ，イロ，イヨ」だとか，形容詞は「カロ，カッ，ク，

イ，イ，ケレ」，形容動詞は「ダロ，ダッ，デ，ニ，ダ，ナ，ナラ」などと習う。だから，語尾がどういうひらがなかがわかれば，動詞か，形容詞か，形容動詞か解かる。また，助詞，助動詞は，かな書きで，1～3文字であることを知る。名詞は経験で漢字書きされることが多い。漢字書きの語が出てくれば大概名詞とする。わからないものは副詞としたものだ。だから連体詞の存在に気づいたのは中学も高学年になってからであった。

この方法も計算機にのせることができる。入力された語がどういう文字で書かれているかを調べておき，語尾が漢字，カタカナなどであれば名詞，「カロ」であれば形容詞の未然形，「イ」であれば，形容詞の終止・連体形か動詞の未然・連用形である，などとする。助詞，助動詞は全部あげても200に満たないだろうから，全部記憶しておき，その語と一致しないかを調べればよい。

3）接続による方法

文法の時間，助詞，助動詞にはいると，接続について教わる。試験に「このぬは何か」という問題がでてくる。直前の語が動詞の連用形なら「完了を現わす助動詞ぬの終止形」と答えねばならないし，直前の語が動詞の未然形ならば「否定の意を現わす助動詞ずの連体形」と答えねばならない。そのどちらでもない場合は，何かの語の一部分かと疑ってかからねばならない。また，連体形は直後に体言がくるとか，助詞「が」を接続するのは名詞しかないということなども知る。

この方法も計算機にのせることができる。まず今言ったような知識を書き込んだテーブルを作る。たとえば「セル」は直前に動詞の未然形がくるとか，「ノ」は，直前に「ト，カラ，デ，ヘ，ヨリ……」などの助詞がくることがあり，活用語の終止形がくるとか，そうでなければ直前は名詞だと書かれている。このようなテーブルをすべての助詞，助動詞，および品詞，記号について作っておく。次に計算機は読み込まれた文の最後の語を捜し出し，それが今，用意したテーブルにあるか調べ，あればその情報を付け，直前の語を予想する。そして次の語に移る，という方法である。

3. プログラム説明

方法の概略で述べた方法に従ってプログラミングした。プログラム名称は，辞書による方法のプログラムをＦＫ１，語形による方法のプログラムをＦＫ２，接続による方法のプログラムをＦＫ３と付けた。ＦＫ２はＦＫ１の，ＦＫ３はＦＫ１と２の手法を用いている。又，ＦＫ３の入力データはＦＫ２の出力結果を用いる。ＦＫ１は辞書の関係で短単位の品詞しか付けられないし，ＦＫ２と３は文節から助詞，助動詞を除いた部分の品詞を付けるように設計されている。後に掲げたフローチャートを参照のこと。

① 辞書による方法（ＦＫ１）

出力形式をどうするかによって二つの方法がある。一つは入力文の形をそのままにして，情報を付け，出力する方法（プログラム名称をＦＫ１—１とする）。他の一つは辞書の配列順に単語を並べ換えて，情報を付け，出力する方法（プログラム名称ＦＫ１—２とする）である。

ともに辞書は新聞の語彙調査の結果を用いる。従って，付けられる情報は短単位についてである。又，辞書にない語は情報が付けられない。

ＦＫ１—１　（フローチャートは116ページ）

入力原文の各語について一つずつ辞書を引き直して行く方法である。辞書を引く時なるべく速くその語に当たるように辞書の語を配列しなければならない。その方法はいろいろあるが，最も簡単なものとしては，辞書の語を度数順に並べておく方法がある。

（説明）

単位切り済みの原文を読み込む。読み込んだ最初の語を取り出し，その語が辞書にあるかを調べる。辞書になければ「不明」の情報を与え，あれば，辞書の情報を転写する。次の語も同様にして情報を付ける。すべてに付け終われば出力する。出力結果は次の通りである。

（辞書フォーマット）

	語	情報	区切り記号	

（出力結果例(注5)）

連休（T100） はじめ（S200／ＳＥＩＤ） に（S100／ＳＥG9／WR00） お（S600） 伊勢（W800） 参り（S200／ＳＥＦＦ） を（WR00） し（ＳＥK5／WR00） て（S100／WR00） 来（ＳＥK3） た（S100／ＷＰＰ0） 。（ＹＹ00）……

ＦＫ１－２ （フローチャートは117ページ）

辞書により情報を転写し終わればすぐアウトプットに移るから，辞書の配列は，その出力の配列順序と同じであれば便利である。五十音順がよい。

（説明）

単位切り済みの入力原文を読み込む。すべてのデータを読み終わったところで，それを辞書の配列と同じ配列順にソートする。ソートすれば同表記語は同じ所に集まるから，それを一つの見出しに直し，度数を付ける。次に情報付けに移る。ソートされた最初のデータと辞書の最初の語を比べる。同じ語であれば辞書の情報を転写する。同じでなく，データの方が配列順序が下位ならば辞書にそのデータと同じ語がないのであるから，「不明」の情報をつける。このようにしてすべて情報を付け終わればそのままアウトプットすればよい。処理結果は次の通りである。

（辞書フォーマット）

	語	情報	区切り記号	

（処理結果例(注6)）

伊勢（W800）	1	し（ＳＥK5／WR00）	1
お（S600）	1	た（S100／ＷＰＰ0）	1
来（ＳＥK3）	1	て（S100／WR00）	1
最近（T100）	1	でき（ＳＥK3）	1

泊まっ（ＳＥＦＤ）	1	参り（Ｓ200／ＳＥＦＦ）	1
に（ＳＥＧ9／Ｓ100／ＷＲ00）	1	町はずれ（Ｓ100）	1
の（Ｓ100／ＷＲ00）	1	旅館（Ｔ100）	1
はじめ（Ｓ200／ＳＥＩＤ）	1	連休（Ｔ100）	1
二見（Ｗ800）	1	を（ＷＲ00）	1

② 語形による方法（ＦＫ２）（フローチャートは120ページ）

情報の記号は表１を参照。テーブルには，ＦＫ２の論理では正しい情報が付かない語（104語）と，助詞，助動詞（121語）が登録されている。特殊語のテーブルは三種類（漢字書きの語，漢字まじりの語，ひらがな書きの語）に分かれている。字種判定の結果を用いると，早く目的の語に当たるからである。テーブルは紙テープで入力される。テーブルに登録されている語を増やすのが簡単であるためである。又，見出し語の長さは４文字になっている。これは，新聞語彙調査一年分の長単位調査結果（度数６以上 11044 異り語数）について調べたところ，５文字以上のもので，ＦＫ２の論理で処理できない語がなかったためである。実験の結果５文字以上で正しく付かない語があればこの長さを増やせばよい。

表1　コード表

1　桁　目		2　桁　目	
コード	品　　詞	コード	活　用　形
0	名　　詞	8	未　然　形
H	動　　詞	9	連　用　形
I	形　容　詞	＃(C)	未然連用形
＋	形容動詞	H	終　止　形
A	連　体　詞	I	連　体　形
B	副　　詞	＋	終止連体形
C	感　動　詞	Q	仮　定　形
J	接　続　詞	R	命　令　形
N	助　　詞		
R	助　動　詞		
P	助・助動詞		
X	記　　号		

テーブルフォーマット

	語	情報	区切り記号	

テーブル例

助詞・助動詞のテーブル　121語

なかっ	R9／ながら	N　／なかろ	R8／なく	R9／なけれ	RQ／
など	N　／なら	RQ／なり	N　／に	N　／ぬ	R＋／
ね	PQ／の	N　／ので	N　／のに	N　／は	N　／
ば	N　／ばかり	N　／へ	N　／ほど	N　／まい	R＋／
まし	R9／ましょ	R8／ます	R＋／ますれ	RQ／ませ	R＃／
まで	N　／も	N　／や	N　／やら	N　／よ	N　／
よう	R＋／ようだ	RH／ようだっ	R9／ようだろ	R8／ようで	R9／

特殊語のテーブル（漢字書きの語3語，漢字まじりの語10語，ひらがな書きの語91語）

得	H＃／又	J	／来	H9／EF					
下さい	HR／一つ	0	／同じ	B	／子ども	0	／特に	B	／
再び	B　／二つ	0	／最も	B	／初めて	B	／少し	B	／
ため	0　／つい	B	／さる	A3／うえ		0H／つぎ		0H／	
こと	0　／もの	0	／その	A	／この	A	／ほか	0	／
また	BA／それ	0	／とき	0	／あと	0	／うち	0	／

（説明）

このプログラムは大きく二つの部分に分かれる。一つは前処理的なもので，語を構成している文字の字種判定である。この結果により，どのテーブルをひけばよいか，語形判定のルーチンに入ってもよいかなどの指示を与える。（この結果を用いて，語種の判定がある程度できるのではないかと考えている。）もう一つは語形判定のルーチンである。これがこのプログラムの基本となるところである。

まずテーブルを読み込む。これは，テーブルの項で説明した理由で紙テープになっている。又，テーブルの語数が少ないので，外部記憶装置にたくわえる

ということはしない。次に単位切りされた入力原文を読み込む。次は字種判定である。

字種判定＊1　各文字に次のような記号を与え，文字列を記号列に換える。

ひらがな——S，カタカナ——T，漢字——U，英文字——V，数字——W，記号——X

例.「　ガラス　」　は　外来語　だから　U　を　つける　。

X　TTT　X　S　UUU　SSS　V　S　SSS　X

次に語を前から取り出す。字種判定の結果を用いてテーブルを捜す。すべて漢字書きならば，まず漢字書きの語のテーブルを，漢字混りの語であれば漢字混りの語のテーブルを，ひらがな書きの語であれば，まず助詞，助動詞のテーブルを次にひらがな書きの語のテーブルを捜す。次に語末の文字（1～2字）により，情報を付けるルーチンに行く。

語末の文字を調べる＊2

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

矢印の左側の検査をし，その通りなら右側の情報を付ける。そうでなければ，次のチェックに移る。数字は順序性を持つ。20～23のチェックはＦＫ２だけでは用いないで，ＦＫ３との接続の時に用いる。語末の文字の検査で，すべてＮＯだったものは「不明」となる。情報が付けば，次の語に移る。すべての語に情報を付け終われば出力する。出力結果は，処理結果例のように入力原文に情報が付いた形となる。

ＦＫ２　処理結果例

多数　（0　　0）決　（0　　0）の　（N　　4）原理　（0　　0）

多数　（0　　0）決　（0　　0）の　（N　　4）原理　（0　　0）に　（N　　4）は　（N　　4），　（X　　2）確かに　（＋9　　8）相対　（0　　0）主義　（0　　0）的な　（＋I　　9）意味　（0　　0）が　（N　　4）ある（H＋　　A）。　（X　　2）甲論　（0　　0）　乙◆日送　（0　　0）の　（N　　4）意見　（0　　）の　（N　　4）対立　（0　　0）が　（N　　4）ある　（H＋　　A）場合　（0　　0）に　（N　　4），　（X　　2）神　（0　　0）なら　（RQ　　4）ぬ　（R＋　　4）人間　（0　　0）の　（N　　4）知性　（0　　0）を　（N　　4）もっ　（H9　　I）て　（N　　4）して　（?　　L）は　（N　　4），　（X　　2）その　（0　　5）中　（0　　0）の　（N　　4）どれ　（HQ　　B）を　（N

4）選ぶ （H J）べき （? L）か （N 4）を （N 4）絶対 （0 0）の （N 4）確信 （0 0）を （N 4）もっ （H9 I）て （N 4）断定 （0 0）し （N 4）うる （H＋ A）者 （0 0）は （N 4）ない （RH 4）。 （X 2）それ （0 5）を （N 4）， （X 2）なお （B 5）かつ （? L）ひとり （? L）の （N 4）絶対 （0 0）の （N 4）権威 （0 0）を （N 4）もっ （H9 I）て （N 4）断定 （0 0）する （H＋ A）と （N 4）いう （? L）制度 （0 0）は （N 4）， （X 2）独裁 （0 0）主義 （0 0）で （P9 4）ある （＋9 A）。 （X 2）独裁 （0 0）主義 （0 0）に （N 4）走っ （H9 I）て （N 4）， （X 2）人間 （0 0）の （N 4）合理 （0 0）性 （0 0）を

原文は江川清氏の「自動単位分割」のプログラムによって単位切りされたものである。文中「◆日送」は「駁」である。（ ）内は前4ケタが品詞・活用情報（2組）で，最後の1ケタはFK2プログラム中のどの箇所で決定されたかを示す記号である。品詞・活用情報は表1を参照。

③ 接続による方法（FK3）（フローチャートは118・119ページ）

テーブルは二種類に分かれる。一つは助詞，助動詞に関する情報をもったテーブル，もう一つは品詞に関する情報をもったテーブルである。テーブル1の「情報」は品詞，用法，終止形の語形，見出し語の活用である。たとえば「助動・否定・ない・終止」とあるのは，「この語は助動詞で否定の意に用いられる〃ない〃の終止形である。」とよむ。制限情報1は見出し語の直前に何がくるかを現わした情報である。これは二つに分かれる。一つは見出し語の前にくる助詞はどういうものかを現わしたもので，これは語を並べてある。もう一つは品詞活用情報で，見出し語の直前にくる語はどういう品詞，活用であるかを現わしたものである。これは，二桁で一組の品詞・活用情報を現わす。制限情

報2は見出し語の直後にはどういう語がくるかを現わしたものである。現在はこの情報は用いない。

現在用意している語数はテーブル1は121語，テーブル2は14語である。これらは紙テープで作られている。変更が簡単なためである。

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　@　@EI

が@格助・接助@＃の＃と＃から＃まで＃も＃だけ＃ばかり＃こそ＃さえ＃のみ＃など＃ぐらい＃　I　＋　H0　／0　@　@EI

た@助動・過去・た・終止連体@H9　9／H9@　@EI

ない@助動・否定・ない・終止@＃せ＃させ＃れ＃られ＃たがら＃は　H8／H8@　@EI

テーブル2

X　@＃か＃さ＃ぞ＃ね＃よ＃　H　＋@　@EI

T　@＃が＃て＃し＃　9　＃　I　＋J　C　@　@EI

0　@　I　+A　0　@　@EI

H　@#て#ては#ても#　I9I#F9FI@　@EI

（説明）

ＦＫ３の入力原文は，ＦＫ２の出力結果を用いる。ＦＫ３のプログラムに入る前に簡単な品詞情報が付いていることが必要なためである。このプログラムの特徴はテーブルの指示に従って品詞を決めて行くところにある。

プログラムを入力原文例によって追ってみよう。「　だれ　（HQ　　B）も　（N　　4）る　（?　　L）ない　（RH　　4）。　（X　　2）　」

まず，このプログラムは後の語から始める。最初に，最後の句点（。）を捜す。情報を付け，句点の制限情報を取り出す。

X　@#か#さ#ぞ#ね#よ#　　H　+@　@Ei

次に制限情報の最初の語は助詞かを調べる。この場合，助詞「か」であるから，フローチャートの下へ進む。制限情報内の助詞「か」を取り出す。次に，今調べている語「。」の直前の語「ない」を取り出す。「か」と直前の語「ない」を比べる。等しくはないから，次の制限情報「さ」について同様のことを調べる。やはり等しくはない。このようにして「よ」まで行く。次はスペースだから，制限情報１は助詞ではない。@でもないから制限情報１はおわりでもない。制限情報内の品詞，活用情報を取り出す。「H」である。次に今調べている語の直前の語「ない」の情報を取り出す。「RH」である。これらを比べる。この場合，品詞でも活用でもどちらか一方が等しくなければ全体は等しいということにしている。従って「H」で等しくなっている。次に「ない」は助詞，助動詞かを調べる。これは助詞であるからテーブルに「ない」を捜す。下記のようにテーブル内に存在する。

ない@助動・否定・ない・終止@#せ#させ#れ#られ#たがら#は　H8／H@　@EI

で，その情報を付ける。「ない（助動・否定・ない・終止）」。制限情報を取

り出しておいて，次の語「ゐ」に移る。さて，取り出した制限情報1は助詞かを調べる。「せ」だから助詞である。これと直前の語「ゐ」とを比べる。等しくないから次の助詞に移る。助詞「は」まで等しくはない。次に，品詞，活用情報と比べることになる。「ゐ」の品詞情報「？」と制限情報「Ｈ８」とを比べると一致しない。次の制限情報は「／」である。これは強制入力で次にある情報を強制的に付けてしまう。従って「ゐ（Ｈ８）」となる。次に「Ｈ」の制限情報を取り出す。

Ｈ　@#て#ては#ても#Ｉ９Ｉ#Ｆ９ＦＩ@　@ＥＩ

次の語に移る。同様に，制限情報の最初の語「て」と「も」を比べる。これは「ても」まで等しくない。次に品詞，活用情報を調べる。「Ｎ」と等しくなるものはこの制限情報の中にはない。しかも強制入力もないから，制限情報おわりに入る。で，今調べている語「も」がテーブルの中にあるかどうかを調べてみる。

も@副助@#に#を#と#から#で#へ#より#まで#の#さえ#すら#など#ぐらい#　９　#　Ｉ　＋０　@　@ＥＩ

あるからこの情報を付ける。「も（副助）」。次の語に移る。同様に「だれ」と制限情報内の助詞と比べる。すべて等しくない。次に，「ＨＱ」と制限情報内の品詞，活用情報を比べる。これもどれも等しくないので強制入力により，「0」の情報が付けられる。「だれ（0）」。

このようにして，処理結果「だれ（0）　も（副詞）　ゐ（Ｈ８）　ない（助動・否定・ない・終止）　。（Ｘ）　」となる。

ＦＫ３　処理結果例

原文

ある日の暮れ方の事である。一人の下人が，◆久公生門の下で雨やみを待ってゐた。広い門の下には，このをとこの外にだれもゐない。

入力原文（ＦＫ２済）

ある　（Ｈ＋　Ａ）日　（０　０）の　（Ｎ　４）暮れ方　（０　０）の　（Ｎ　４）事　（０　０）で　（Ｐ９　４）ある　（Ｈ＋　Ａ）。　（Ｘ　２）一人　（０　０）の　（Ｎ　４）下人　（０　０）が　（Ｎ　４），（Ｘ　２）◆久公生門　（０　０）の　（Ｎ　４）下　（０　０）で　（Ｐ９　４）雨やみ　（？　Ｌ）を　（Ｎ　４）待っ　（Ｈ９　Ｉ）て　（Ｎ　４）る　（？　Ｌ）た　（Ｒ＋　４）。　（Ｘ　２）広い　（Ｉ＋ＨＣ６）門　（０　０）の　（Ｎ　４）下　（０　０）に　（Ｎ　４）は　（Ｎ　４），（Ｘ　２）この　（Ａ　５）をとこ（？　Ｌ）の　（Ｎ　４）外　（０　０）に　（Ｎ　４）だれ（ＨＱ　Ｂ）も　（Ｎ　４）る　（？　Ｌ）ない　（ＲＨ　４）。　（Ｘ　２）

処理結果

ある　（Ｈ＋）日　（０　）の　（格助）暮れ方　（０　）の　（格助）事　（０　）で　（格助・接助）ある　（Ｈ＋）。　（Ｘ　２）一人　（０　）の　（格助）下人　（０　）が　（格助・接助），　（Ｘ　２）◆久公生門（０　）の（格助）下　（０　）で　（格助・接助）雨やみ　（０　）を　（格助）待っ　（Ｈ９）て　（接助）る　（Ｈ９）た　（助動・過去・た・終止連体）。　（Ｘ　２）広い　（Ｉ＋）門　（０　）の　（格助）下　（０　）に　（格助）は　（副助），　（Ｘ　２）この　（Ａ　）をとこ　（０　）の　（格助）外　（０　）に　（格助）だれ　（０　）も　（副助）る　（Ｈ８）ない　（助動・否定・ない・終止）。　（Ｘ　）

注）文中「◆久公」は「羅」である。

4. 各方法の精度，処理時間，問題点

	ＦＫ１	ＦＫ２	ＦＫ３
の	Ｓ100／ＷＲ00	（助詞）	格助
で	Ｓ200／ＳＥＩ８／ＷＲ00	Ｐ９（助詞・助動詞・連用形）	格助・接続助
広い	ＳＬＭＯ	（形容詞・終止連体形，動詞・未然連用形）	（形容詞・終止連体形）

雨やみ	S200	?	0（名詞）
る	SEGG	?	8（動詞・未然形）， 9（動詞連用形）
ある	SEFF／SD00	（動詞・終止連体形）	（動詞・終止連体形）

ＦＫ３に用いた例文に各方法が，どのような情報を付けるかを比べたものが上の表である。これを例にしながら，各方法の問題点，精度，処理時間などを述べてみよう。

① ＦＫ１

上の例に見られるように，辞書にある語はどんな語でも，情報が付く。しかし，辞書にない語は付かない。ここで用いた辞書が短単位の語であったから，短単位にしか付かない。

テーブル内の情報はすべて付く。従ってＦＫ２や３では付かない語種情報まで付けることができる。しかし，たとえば「の」の情報のように，助詞だけでよいのに，名詞まで付いてしまう。正しい情報だけを取り出すのには，そこでチェックが必要になる。そのチェックのプログラムのひとつとしてＦＫ３を用いることができる。

ここで用いた辞書は，前にも述べたように新聞語彙調査短単位一年分の調査結果である（47805異り語数）。

語彙調査として致命傷であるが，同形異語の判別ができない。ＦＫ１－１は原文の形を残しているので，あとで修正が可能である。しかし，処理速度は遅くなる。

② ＦＫ２

この方法は辞書による方法のどんな語でも情報を付けることができるという長所を取り，辞書の語数を減らすことにより，辞書引きの時間を短縮するという短所を解消している。

語形により情報を付けているので，同形異語の判別はできない。上の例の「で」や「広い」がその例である。又，付け間違うこともある。それは，辞書にある語形に同形異語が存在する時である。たとえば，「の」の名詞や，「し」の動詞などがそれである。

このプログラムはわからない語が出てくれば，テーブルを補充する形式を取っている。

情報の付き方に四種類ある。一つは正しく付く場合。二つめは完全な間違い。三つめは不必要な情報が付いている場合。四つめは不明の場合である。語末のチェック23までを生かすとすれば，新聞語彙調査一年分長単位語で，度数6以上の語11044異り語で，三つめの種類の間違いは20語である。これは，このデータを用いてこの論理を作ったためであるが，かなりの正解率を示すことが解かる。語彙調査に用いるプログラムとして，二つめの種類の間違いが一番困る。この場合だけ，どこに間違いがあるか解からないからである。

327語の文章（多数決の原理，羅生門の一節）の実験では90.2%の正解率を示す。この中には三つめの種類の間違いも入っているが，これを除くと，83.2%の正解率となる。処理時間は，250語の文で1分50秒。入出力に（紙テープ）50秒かかっている。これを除けば1秒間に4～5語処理することになる。

③ ＦＫ3

このプログラムはＦＫ2のアウトプットを用いる。

従って，このプログラムは辞書方式のどんな語でも付くという点，語形方式の処理時間が短く，辞書にない語も付くという点，接続方式のその文における働きによって付けるという点のすべての長所を取ったプログラムといえる。

このプログラムは，同形異語の判別を行なう。上の例の「で」がそれである。

全く未知の語でも助詞，助動詞が付いていればかなりの情報を付ける。上の例の「雨やみ」や「ゐ」がそれである。

かなり正確な活用情報を付ける。「ゐ」の末然形，連用形の判別がそれである。

精度は327語の文で94.8%，二つめの種類の間違いを除くと92.3%になる。

現在のプログラムではＦＫ2のプログラムよりかなり遅い。これは，ＦＫ2と完全に組み合わさっていないためである。

5. おわりに

語彙調査には同形異語の判別は欠かせない条件となる。ＦＫ３では少し行なうものもあるが，なお完全でない。これは，ＦＫ３の方式をより詳しくするとともに新しい方式（狭い範囲での構文の解析など）を考えなければならないだろう。たとえば「カキ」という語では，動詞の「書く」の連用形と，名詞の「カキ」とはＦＫ３で判別できるが，木になる柿か，海で取れる牡蠣かは構文の解析の方法を考えなければならない。「カキを酢で和える」と言えば海の牡蠣だろうし，「カキの皮をむく」と言えば木になる柿だろう。構文の解析にしろ，接続にしろ，用例を集めての研究が必要である。

今後の方向としては，ＦＫ１～３のプログラムの統合，精度を高めること，及び同形異語判別ルーチンを組み込むことを考えていく予定である。

（この報告は，昭和45年６月１日国立国語研究所の研究報告会で発表したものに加筆したものである。）

（注１）　斎藤秀紀「電子計算機による語彙調査――主として長単位処理について――」国立国語研究所報告34「電子計算機による国語研究Ⅱ」，「電子計算機による語彙調査Ⅱ――主として短単位処理について――本報告に掲載。
国立国語研究所報告37「電子計算機による新聞の語彙調査」５ページ。

（注２）　第一資料研究室「語彙調査データの一貫処理法の研究」ＬＤＰ月報別冊４。

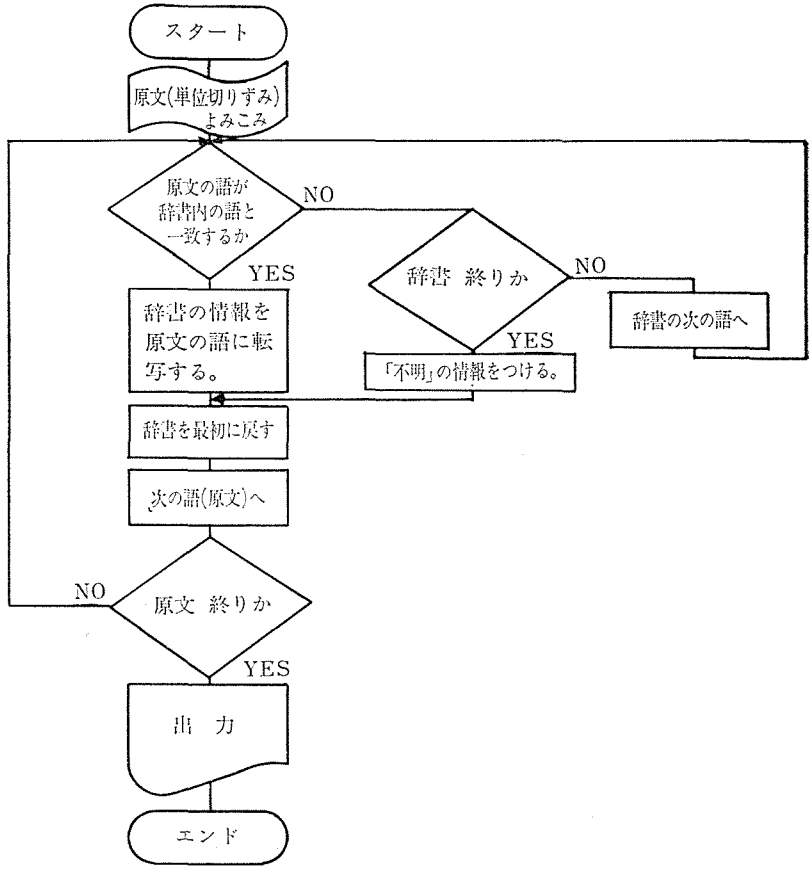
（注３）　江川清「単位分割自動化のシステムについて」計量国語学51。

（注４）　田中章夫「漢字の自動解読システムについて」計量国語学48。

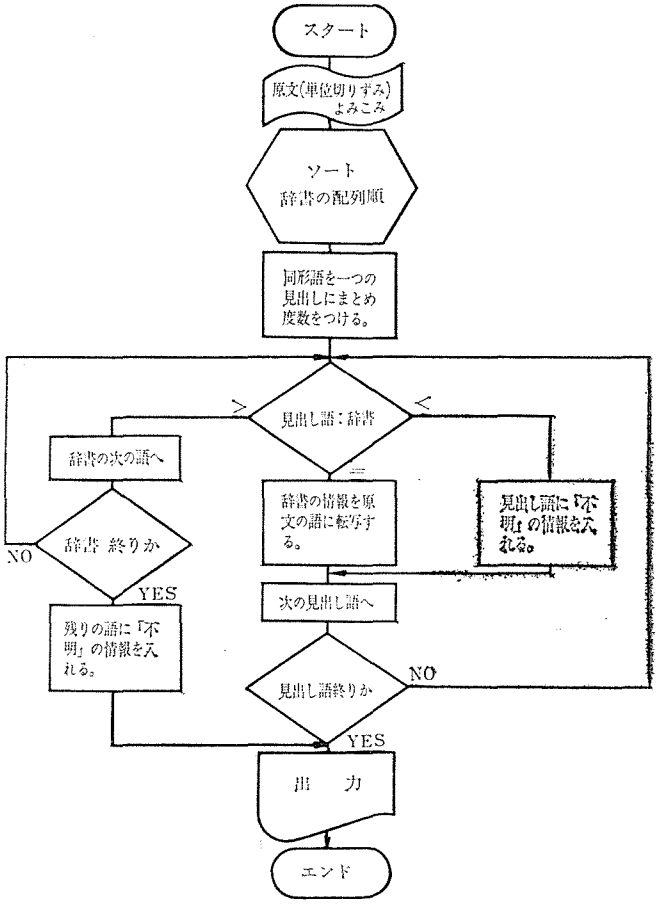
（注５）　石綿敏雄「電子計算機による語彙調査の一実験」国立国語研究所論集２。

（注６）　中野洋「新聞語彙調査の類別語彙表について」国立国語研究所報告34　52ページ。

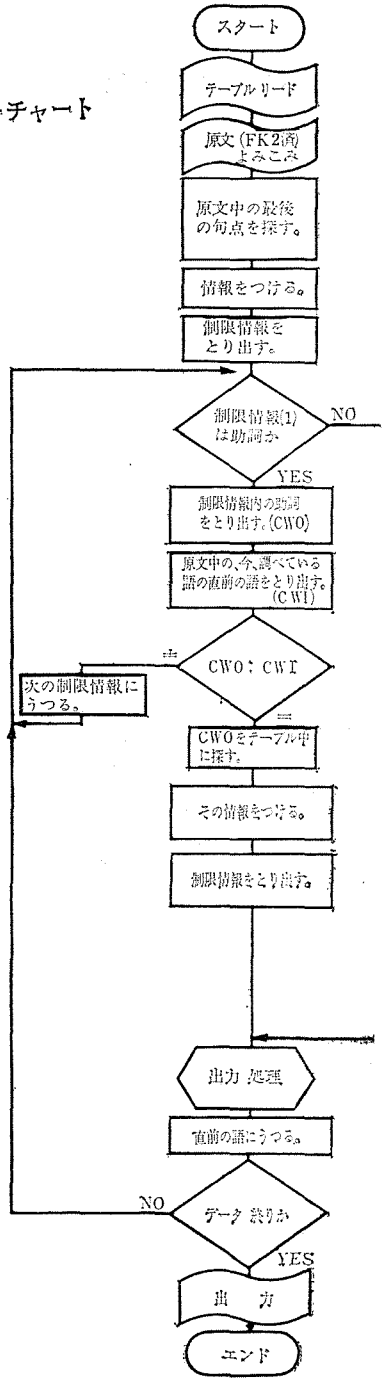
FKI-1 フローチャート

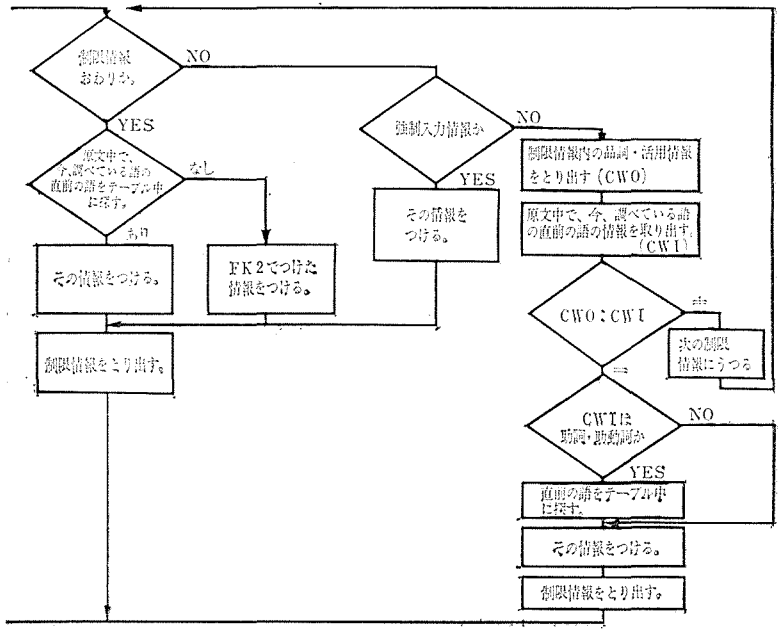


FK1-2 フローチャート

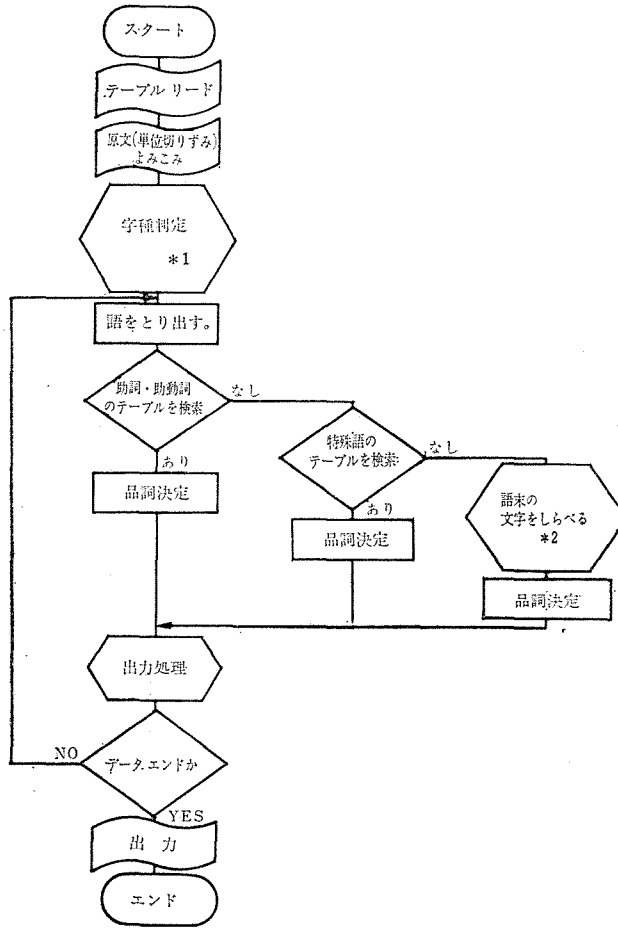


FK3 フローチャート


スタート
テーブルリード
原文(FK2商)よみこみ
原文中の最後の句点を探す。
情報をつける。
制限情報をとり出す。
制限情報(1)は助詞か
NO
YES
制限情報内の助詞をとり出す。(CWO)
原文中の、今、調べている語の直前の語をとり出す。(CWI)
CWO : CWI
≠
次の制限情報にうつる。
=
CWOをテーブル中に探す。
その情報をつける。
制限情報をとり出す。
出力処理
直前の語にうつる。
NO
データ終りか
YES
出力
エンド


制限情報
おありか。
NO
YES
原文中で、今、調べている語の
直前の語をテーブル中
に探す。
なし
あり
その情報をつける。
FK2でつけた
情報をつける。
制限情報をとり出す。
強制入力情報か
NO
YES
その情報を
つける。
制限情報内の品詞・活用情報
をとり出す（CWO）
原文中で、今、調べている語
の直前の語の情報を取り出す。
（CWI）
CWO：CWI
≠
次の制限
情報にうつる
=
CWIは
助詞・助動詞か
NO
YES
直前の語をテーブル中
に探す。
その情報をつける。
制限情報をとり出す。

FK2 フローチャート


スタート
テーブル リード
原文(単位切りずみ)よみこみ
字種判定 *1
語をとり出す。
助詞・助動詞のテーブルを検索
なし
あり
特殊語のテーブルを検索
なし
あり
品詞決定
品詞決定
語末の文字をしらべる *2
品詞決定
出力処理
NO
データ エンドか
YES
出 力
エンド