Automatic processing of Japanese sentence for word counting by personal computer

国立国語研究所報告 103　研究報告集12（1991）

パソコンによる語の認定処理

中野　洋

NAKANO Hiroshi : Automatic Processing of Japanese Sentence for Word Counting by Personal Computer

要旨：
(1) 語彙調査支援のための自動処理プログラムシステムの中核である一貫処理プログラムを作成し，これをパーソナルコンピュータに移植した。
(2) 一貫処理の機能は，単語分割，読み仮名付け，品詞認定，語種認定，活用形変換である。
(3) このプログラムの特徴は，プログラムと辞書が小さいこと，どのような文でも処理できること，処理が早いことである。プログラムはC言語で書いた。MS-DOSで128Kバイトの容量があるパソコンであれば動く。
(4) 語彙調査データの作成作業における人手の作業と機械処理の比較を行った。その結果，次の3点が明らかになった。①処理精度は，単位切りでは機械でほぼ90%，人手では97%～98%が見込まれることがわかった。これは明らかに人手の方がよい。②処理時間は，機械は人手の10%以下である。③入力パンチ量については，機械は人手の約20%である。
(5) 一貫処理プログラムは，処理方法とともに辞書が重要である。付録として主要な辞書を示した。

キーワード：単位分割，品詞認定，漢字解読，活用形変換，語種認定，パーソナルコンピュータ

Abstract : This paper describes a Japanese sentence analyzing program executed by personal computer.
This program has the following functions:
1. segmentaton of Japanese sentences
2. transliteration of Chinese characters into the Japanese syllabary
3. classification of the parts of speech and the word origin of the Japanese vocabulary
4. transformation of the conjugation forms of the verbs into the dictionary form
The accuracy of the program is 90%, but when performed by a human subject, it reaches 97−98 %. The time needed by a human subject however is ten times that of the computer by itself, while the amount of the input data takes five times longer.
The program can be said to be useful for word counting.

Key words : segmentation, classification of the parts of speech, classification of the word origin, transliteration of Chinese characters into the Japanese syllabary, transformation of the conjugation forms, parsonal computer.

１．はじめに

国立国語研究所では，昭和41年に電子計算機HITAC-3010を導入し，新聞３紙の語彙調査を実施した。延べ語数は約300万短単位である。それ以降，「漱石・鷗外の用語調査」，「高校教科書の語彙調査」「中学校教科書の語彙調査」を，同様の汎用電子計算機で行ってきた。電子計算機の大きさは，それぞれの時代でほぼ中型機に分類されるものである。

一方，パーソナルコンピュータは，昭和55年にPC-8001を導入したが，これは大量の語彙調査にはまだ適さないものであった。漢字処理ができない，外部記憶装置に高速大量の媒体がないという理由である。しかし，操作性は大型コンピュータに比べ優れていたので，言語研究のためのデータベースを操作するプログラム（DBMS）を開発し（文献28），これを大型計算機に移植した（文献26）。ただし，前述の理由でカタカナデータ（話しことばデータ）に適用したものであった。

現在，パーソナルコンピュータは，その処理速度や外部記憶装置の容量が改良され，各種の汎用プログラムが開発されている。特に，言語研究には処理速度より，操作性が重要であり，その点でもパソコンは言語研究に適していると考えられる。

入力装置としてのパソコンは，ワープロの普及に見られるように個人用としては最適のものである。大容量の外部記憶装置が出現し，大量データの処理も可能になりつつある。また，出力装置としての熱転写プリンタやドットプリンタのスピードは大量データの印字には適していないが，最近のレーザービームプリンタの速度・印字品質は大量処理にも使えるレベルにまで向上しているといえる。ワークステーションは，パソコンの機能をさらに大きくした研究用機器として利用されているが，そのソフト環境や能力は十分にこの種の研究に耐えるものである。

最近の言語研究の研究環境は，このような電子機器の発達により大きく変りつつある。その第１は，研究の対象である日本語が機械可読形式になりつつあることである。たとえば，新聞や雑誌，単行本のおおくは，電算写植に

よるものとなりつつある。これは電子計算機で処理できる形である。その第2は，入力機械としての光学漢字読取り装置の開発である。印刷されたものをそのまま電子計算機に入力することが出来る。その第3は，データベースの構築が行われつつあることである。国立国語研究所と国語学会が共同で発行した「日本語研究文献目録・雑誌編」[フロッピー版]（文献8），各種のフロッピー化された索引類（文献18），ＣＤ－ＲＯＭ媒体など（文献4，15，19,）がそれである。これらはすべて日本語研究の対象となりえる。

以上の日本語データの多くは，通常の表記体である分ち書きしない漢字仮名混じり文である。それらは，日本語処理（日本語で書かれている情報を処理すること）にも日本語研究にも用いることができるが，その多くはまずもとのデータを単語に分割するところから始めなければならない。

国語研究所のこれまでの語彙調査では，これを人手で行ってきた。しかし，その労力，費用，人手は膨大なものであり，その省力化が望まれていたのである。

以下に述べる一貫処理法の開発はそれにこたえるものであった。

２．目的

国立国語研究所では，語彙調査支援のための自動処理プログラムの開発を行ってきた。

その一つの成果が一貫処理プログラムを中心とした語彙調査支援システムである。

このシステムは，次の５つのサブシステムからなる。

(1) 一貫処理
　単語分割，読み仮名付け，品詞認定，
　語種認定，活用形変換，
(2) KWIC作成システム
(3) 修正・同語異語の判別システム
(4) 語彙表作成システム

(5) 集計・分析システム

このうち，(2)～(5)のシステムは以下の通りである。

「KWIC作成システム」の主なプログラムは，用例付け・ソートからなり筆者もすでに報告し（文献21），またすでに多くの人が独自にプログラムを作成し利用している。

「修正・同語異語の判別システム」は，原データを修正したり，KWICデータや単語データに新たな情報を付けたりするプログラムシステムである。大量データの処理には欠くことのできない処理であるが，パソコンで使われている各種のエディタが色々な機能を持っていて，便利である。

「語彙表作成システム」は，集計データを表の形に作成印字するプログラムの集まりである。これらは，主にワープロソフトの印字機能を用いる方が便利である。

「集計・分析システム」は，調査の目的にあわせたプログラムが必要である。しかし，その中でもたとえば語数カウントや比率計算，度数ソート，五十音ソート等は各語彙調査に共通のプログラムだろう。これらのプログラミングはそれほど難しくないし，また表計算ソフトなどが市販されており，利用することができる。

以上のプログラムシステムは各語彙調査において開発し，「電子計算機による新聞の語彙調査」（文献９）や「高校教科書の語彙調査」（文献10），「中学校教科書の語彙調査」（文献11）などの報告書で報告している。詳細は文献を参照されたい。また，これらの機能をもったパソコンでのプログラムは，特別研究「語彙調査自動化のための基礎的研究」において開発した。これについては，機会を得て報告したい。

ここでは，(1)の一貫処理プログラムについて報告する。

３．一貫処理の機能

(1) 一貫処理法の開発の歴史

一貫処理は，電子計算機による語彙調査の手作業部分の自動化にある。人

間による作業は，一般に機械処理の結果と比べ精度は良いが，作業時間・作業人数・費用が多くかかる。また作業ミスも散見され，その現れる箇所が一定しない。これにくらべ，機械処理の結果は，その精度は劣るものの，作業時間・費用が少なく，処理ミスの現れる箇所が一定している。

そこで，国立国語研究所では電子計算機を導入して以来，単位分割・漢字解読・品詞認定の自動化プログラムの開発研究を行ってきた。とくに，国語研究所がそれまで蓄えてきた大量の用語用字調査の成果がその開発に大きく役立った。

これらの３つの自動化プログラムは，昭和43～45年頃に相次いで完成した。しかし，その統合については開発が遅れ，完成したのは昭和55年である。さらに，このプログラムシステムが非常に小型であり，また最近パソコンの機能が格段に向上したので，パソコンへの移植を試みた。この時，同語異語判別のために活用形変換と，語彙分析のために語種認定のプログラムを開発し，機能を追加した。これらは，各種の調査や研究に役立つことが確かめられ，その公開が求められている（文献２，文献14)。

(2) 一貫処理の機能

一貫処理の機能は，単語分割・読み仮名付け・品詞認定・語種認定および活用形変換の自動処理である。前の３者は大型計算機の上で開発しパソコンに移植したものであり，後の２者はパソコンの上で開発したものである。

(3) 一貫処理の特徴

一貫処理は，語彙調査を助けるために開発した。筆者は，語彙調査の完全自動化は望めないと考えている。なぜなら，どのような語が現れるかを調査するのが語彙調査の目的であるが，完全な自動化は完全な辞書と文法がなければ不可能であり，この両者は矛盾するからである。

プログラムには，大きく二つの手法がある。ひとつは，大きな辞書によるもの（辞書方式）であり，ひとつはルールによるもの（プログラム方式）である。前者は，精度が良いが時間がかかる。後者は，精度が落ちるが処理時間が少なくてすむ。

一貫処理は，後者の立場によるプログラムであり，非常に小さい辞書と小さなプログラムによって動く。どのような文章にも適応でき，精度は90％以上を目指している。また，処理ミスの修正は修正システムや同語異語判別工程によることを想定している。

一般に工学系では精度90％程度では実用には堪えないと考えられている。こう考えるのは，検査・修正なしで実用化をはかろうとするからである。言語研究の場合，データを見ない研究は考えられないから，検査・修正は当然のことである。したがって，精度もさることながら手近のパソコンで処理するためのプログラムや辞書の小型化が重要である。

同様な立場をとった単位分割のプログラムに坂本義行氏作成のものがある（文献13)。一貫処理との違いは，これが文節単位の分割であることである。大型計算機での実験では，特許公報を対象として97.5% の精度をあげている。一貫処理の単位分割にくらべかなり精度が良いが，単位の違い（文節単位では，活用語の語尾と助詞・助動詞連続の分割が必要なくなる）や処理対象の違いが考えられる。

4．処理の方法

(1) 漢字仮名変換（漢字解読）

ここでは漢字にその読みを付ける処理を行う。

語彙調査の結果である語彙表では，単語を五十音順に並べなければならない。したがって，語彙調査を機械化するにはプログラムに単語の読みを付ける機能が必要となる。

漢字に読みを付ける方法には大きく２種類がある。一つは，単語の辞書を用いる方法である。他の一つは漢字の字書を用いる方法である。これらにはそれぞれ長所と短所がある。

前者においては，数万語の単語辞書を持たなければ多くの単語に読みが付かない。また，いくら辞書を大きくしても未知語（辞書に無い語）が現れる。辞書が短い単位で構成されている場合，複合語の解析も問題になる。これら

は前者の方法の短所といえる。長所は，プログラムや辞書の作成が簡単なことである。また，読みの難しい語も辞書にさえ登録しておけば正しく仮名を付けることが出来る。

後者においては，プログラムや辞書の作成が難しいことが短所と言えよう。また，連濁や連声，特殊な読み方などの処理に難点がみられる。長所は，字書のオーダーが数千ですむことである。また，字書にある範囲ではどのような漢字にも仮名を付けることが出来る。それだけパソコンにのせやすい。たとえば，目の不自由な人のための文章の読み上げ機械が研究されているが，このような場合，たとえ間違っていても仮名が付かなければ役に立たないのである。

以上の方法の中で，一貫処理では後者の方法をとった。すなわち，漢字の字書を作り，そこから適当な読みを選択する方法である。理由は，その長所を重視したためである。

漢字の読みを選択する方法は次の通りである。すなわち，入力文における漢字の前後の文字環境による。漢字の読みは小さなテーブルに書かれている。このテーブルは１・２・３グループの３種類に分かれている。

グループ１の漢字は１つの読みしか持たない。だから，プログラムは，この漢字が来たらその読みに置き換えるだけでよい。図１の例１の漢字はこのグループである。ここに属する漢字の数は，院・堂・族・宇・批など1240字である。

グループ２の漢字は２つまたは３つの読みを持っている。図１の例２・３はこのグループの漢字である。ここに属する漢字の数は793字である。

テーブルのフォーマットは次の通りである。

グループ番号	漢字	演算記号	読み（４文字まで）
2	歌	1	カ
		A	うた

読みは，表１と入力文における漢字の文字環境によって選ぶ。

表1　環境演算テーブル

環境		演算用コード															
直前	直後	A	1	B	2	C	3	D	4	E	5	F	6	G	7	H	8
非漢字	非漢字	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
非漢字	漢字	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1
漢字	非漢字	1	0	1	0	0	1	0	1	1	0	1	0	0	1	0	1
漢字	漢字	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1

0：漢字をテーブルの中の読みに代える

(1)　1校コウ　　☆
(2)　2歌1カ　　Aうた　　☆
(3)　2河1カ　　Aかわ　　☆
(4)　3川18セン　　2Hかわ　　＊M河1N柳1☆
(5)　3泳11エイ　　2Aおよ　　＊M平2Nぎ2☆
(6)　3水11スイ　　2Aみず　　＊M大2み気2☆

図1　漢字解読辞書

実験文1		実験文2		実験文3	
①	②	①	②	①	②
校	コウ	川	かわ	河	カ
歌	カ	で		川	セン
を		泳	およ	で	
歌	うた	ぐ		水	スイ
う		。		泳	エイ
。				を	
				す	
				る	
				。	

①入力文字列　②漢字の読み

図2　漢字解読実験結果

図２は，実験の結果例である。３つの実験文には，「校・歌・川・河・泳・水」の６つの漢字が用いられている。その中で，「歌・川・泳」は２度用いられ，それぞれ異なる読み方をしている。この選択の方法を以下に述べる。

実験文１の「歌」はグループ２の漢字で，その環境「校歌を」では，前が漢字「校」で後が仮名「を」である。文脈が「漢字＋非漢字」の時には，環境演算テーブルは，漢字解読辞書の「歌」のＡと１の読みのうち１の読み（すなわち，「カ」）を選ぶよう指示している。また，文脈「を歌う」の「歌」の環境は「非漢字＋非漢字」だから，同様にして漢字解読辞書のＡの読み（「うた」）を選ぶ。このようにして，「校歌を歌う」は「コウカをうたう」に変換される。

グループ３の漢字はグループ２の読みの他に特別な文脈における特別な読みを持っている。図１の番号４・５・６がこのグループの漢字である。図２の中の，２番目の文の「川」は特別な処理がされないで読みが与えられる。しかし，３番目の文は特別処理が必要である。すなわち，記号「＊」の後の指定環境が適用される。これは特別な文脈において特別な読み（環境演算テーブルでは与えられない読み）を与える処理（指定演算回路の処理）である。すなわち，テーブルによれば漢字「川」の前（テーブルではM，後ろの場合にはN）が「河」であるか又は後が「柳」であれば無条件に読み番号１の「セン」を与える。このグループの漢字解読辞書のフォーマットは次の通りである。図１の番号４を例にとると，

グループ番号	漢字	読み番号	演算記号	読み（4文字）	記号	前または後の記号	漢字	適用する読み番号
3	川	1	8	セン	＊	M	河	1
		2	H	かわ		N	柳	1

このグループに属する漢字の数はおよそ９１２字である。

下に示す連濁や連声などの現象はグループ３の指定演算回路で処理しなければならない。

「本箱」を「ほん」＋「はこ」ではなく，「ほん」＋「ばこ」とする
「天皇」を「てん」＋「おう」ではなく，「てん」＋「のう」とする
「因縁」を「いん」＋「えん」ではなく，「いん」＋「ねん」とする
「酒屋」を「さけ」＋「や」　ではなく，「さか」＋「や」　とする

(2)　自動単位分割

日本語では漢字仮名まじり文は分かち書きはしない。これをある語の単位に分割することを行う。これにも方法は大きく分けて２種類ある。単語辞書を使う場合と表記の違いを利用する方法である。前者では，大きな辞書が必要なこと，辞書の表記と処理文の表記が合わない場合の処理，未知語が処理できないことが欠点であり，プログラムの作成が簡単なことが長所である。後者は，辞書が無くても良いが，同じ表記が続く場合の処理が難しいことが欠点である。また，日本語の語の単位は大きく短い単位と長い単位に分けられる。ここで用いるのは後者の方法で，長い単位に分割する。

日本語の文章では文字の使い分けをする。図３は新聞の文字の割合を示している。

漢　字	ひらがな	カタカナ		数　字	記号
43.4%	28.0%	8.1%		9.8 %	9.2%

0.6%　（ローマ字）

延べ字数　　1,489,175

図３　新聞における文字の使用分布

日本語文を文字の連続とみて，入力文を次のように字種の列に変えることができる。

Ａ Ｍ ． １ ０ に バ ス に 乗 る 。
英 英 記 数 数 平 片 片 平 漢 平 記

ところで，作文教育においては，文字の使い方を次のように教えることがある。

漢字　　意味を表わす。名詞や動詞の語幹に用いる。

平仮名　助詞・助動詞・動詞の語尾・形容詞・発音にそった表記に用いる。

片仮名　外来語・外国の人名・地名・擬声語・擬態語に用いる。

英文字　外国語の表示・略語に用いる。

数字　　数の表記に用いる。

これらは，異なった文字がそれぞれ単語の種類を表すことを示しているといえる。

単語の切れ目となるところの，文字種連続の組み合わせの頻度を調べたのが表２である。表の割合の高いところで文字列を分割すれば単語に分割することができる。ただし「漢字－平仮名」の連続のところは単語の切れ目となる割合が61.7％と高いが分割しない。というのは，この連続には動詞の語幹と語尾の連続が多数含まれるからである。漢字の直後の平仮名が助詞・助動詞の場合には，後に述べるように図４のテーブルによって分割する。

プログラムでは，表２を表３のように変え，表３の数字の１のところで文字列を分割することにした。

表２．語の切れ目における文字種連続の割合

前＼後	漢字	平仮名	片仮名	英文字	数　字	記　号
漢　字	5.7	61.7	45.2	75.0	100.0	73.8
平仮名	92.1	40.8	95.7	100.0	100.0	95.1
片仮名	25.4	89.5	1.0	——	——	33.3
英文字	2.8	100.0	100.0	13.2	0.0	90.0
数　字	2.7	100.0	——	100.0	0.0	75.0
記　号	98.2	84.7	62.1	33.3	23.7	——

（単位は％，新聞の語彙調査データによる。）

表３．文字連続による単語分割の表

前＼後	漢　字	平仮名	片仮名	英文字	数　字	記　号
漢　字	0	0	0	1	1	1
平仮名	1	0	1	1	1	1
片仮名	0	1	0	0	0	0
英文字	0	1	1	0	0	1
数　字	0	1	0	1	0	1
記　号	1	1	1	0	0	0

0：分割しない　1：分割する

平仮名－平仮名の連続は日本語において最も多い連続である。表２によれば，この連続は分割出来ない。したがって，次の規則を作った。

平仮名「を」は助詞としてのみに使われる。したがって，いつもこの前後で分割する。他の平仮名は図４のテーブルにある文字列をテーブルにしたがって分割する。

字数　文字列（10字以内）①②③①②③①②③

1が	1R
4こうした	2C　1E91P
1た	1P＋
1で	1O9
1の	1R
1れ	1P廾

①：単語の長さ，②：品詞，③：活用

図４．品詞認定・単語分割のためのテーブル例

このテーブルに登録されている文字列は，助詞・助動詞・副詞・表３では分割出来ない文字列などの359である。このテーブルの作成には，斎藤秀紀

氏による漢字仮名混じり文の文字列調査の結果（文献12）を参考にした。

このテーブルは次のように適用される。入力文のなかにテーブルの文字列がないかを探す。もし文字列「こうした」が入力文（例えば「こうした時・・」など）にあれば，テーブルの中の単語の長さによって分割し，品詞や活用の情報を与える。その結果「こう／し／た」のように単語を得ることができる。

文 区切り	文 区切り	文 区切り	文 区切り	文 区切り	文 区切り
C 1 O 1 L 1 I 1 N 1 G 1 8 1 0 1 が 1 東 1 京 1 の 1 都 1 市 1 セ 1	ン 1 タ 1 ー 1 ホ 1 ー 1 ル 1 で 1 開 1 催 1 さ 1 れ 1 た 1 。 1 遊 1 び 1	1 に 1 あ 1 き 1 た 1 子 1 供 1 ら 1 が 1 帰 1 っ 1 て 1 い 1 く 1 。 1 ジ 1	ョ 1 ン 1 ． 1 F 1 ． 1 ケ 1 ネ 1 デ 1 ィ 1 は 1 偉 1 大 1 な 1 大 1 統 1	領 1 だ 1 っ 1 た 1 。 1 パ 1 ン 1 粉 1 を 1 １ 1 ０ 1 ０ 1 ｇ 1 か 1 ， 1	１ 1 ０ 1 ０ 1 円 1 分 1 く 1 だ 1 さ 1 い 1 。 1

区切り欄が 1の箇所で語が切れる。

図５．単位分割実験の結果

図５は単語分割と漢字解読の結果である。表３によって「ＣＯＬＩＮＧ８０」，「東京」，「都市センターホール」，「開催さ」の文字列が得られ，図４のテーブルにより，「が」，「の」，「で」，「れ」，「た」が分割される。

(3) 自動品詞認定

語彙調査における分析の一つとして，品詞分類を行なう。このプログラムでは３つの方法によってこれを実現している。

１番目の方法は図４のテーブルによる方法である。

２番目の方法は，以下に示す規則を用いた語形による方法である。その規則を適用した場合の精度をそれぞれの規則の後に（）付きで示した。

1．もし語末の文字が漢字か，片仮名か英文字であれば，その単語は名詞である。(94.4%)

2．もし語末の文字が「い」であれば，動詞の連用形か，形容詞の終止形または連体形である。(86.2%)

3．もし語末の文字が「く」であれば，動詞の終止形または連体形か，形容詞の連用形である。(83.4%)

4．もし語末の文字が「る」であれば，動詞の終止形である。(95.8%)

5．もし語末の文字が「れ」であれば，動詞の仮定形か，指示代名詞か，助動詞である。(92.9%)

6．もし語末の文字が「ろ」であれば，動詞の命令形か，名詞である。(63.3%)

7．もし語末の2文字が「かっ」であれば，形容詞の未然形か，動詞の連用形である。(74.2%)

8．もし語末の文字が「っ」であれば，動詞の連用形である。(79.6%)

9．もし語末の2文字が「漢字＋平仮名」であれば，それは動詞である。(94.4%)

最後の文字の平仮名の母音が／a／であれば，その語の活用形は未然形または連用形である。

最後の文字の平仮名の母音が／i／であれば，その語の活用形は未然形または連用形である。

最後の文字の平仮名の母音が／u／であれば，その語の活用形は終止形または連体形である。

最後の文字の平仮名の母音が／e／であれば，その語の活用形は仮定形または命令形である。

最後の文字の平仮名の母音が／o／であれば，その語の活用形は命令形である。

10．もし語末の文字が数字であれば，それは数字であり，語末が記号であれば，記号である。

3番目の方法は語の接続のしかたを利用する方法である。すなわち，日本語の文法において語結合－特に名詞や動詞と助詞・助動詞－は自由ではない。その規則によって図6のようなテーブルを作った。

フォーマットは次の通り。(@は区切り記号である。)

①語

②品詞

③この語の直前に用いることのできる助詞・助動詞

④この語の直前に用いることのできる品詞と活用形

⑤もし，直前の語が3・4と一致しなければ強制的に適用する品詞・活用形

```
①  ②  ┌──────────────────────────③──────────────────────────
を@R@#と#から#まで#の#だけ#ばかり#こそ#さえ#すら#のみ
──────────┐  ④     ⑤
#など#ぐらい#  1     /1   @   @
```

図6．品詞接続テーブル

図7は自動品詞認定の実験結果である。

①　②	③	④⑤	⑥⑦
祭　まつ			
り	1	E#	1
を	1	R	R
待　ま			
って	1	E9	E9
い	1	R	R
る	1	E+	E+
。	1	Y	Y

図7．品詞認定実験の結果

①入力文
②漢字解読の結果
③単位分割の結果
④方法1・2による品詞認定の結果
⑤活用形
⑥方法3による品詞認定の結果
⑦活用形

品詞コード
1: 名詞　A: 接続詞 B: 感動詞
C: 副詞　D: 連体詞 E: 動詞
M: 形容詞 P: 助動詞 Q: 助動詞，助詞
R: 助詞　Y: 記号　X: 数字

活用コード
8:未然形 9:連用形 #:未然形，連用形
H:終止形 I:連体形 +:終止形，連体形
Q:仮定形 R:命令形

図7において，④⑤は方法1のテーブルによる品詞認定と方法2の語形による品詞認定の結果である。方法3で修正した結果が⑥⑦である。最初「祭り」は語形が「漢字＋平仮名」の形なので動詞とされた。しかし，方法3の語の接続のしかたによって，つまりこの場合図6のテーブルを用いて助詞「を」の直前には指定の語が来ていないので強制的に名詞に変えられたのである。

(4) スーパーバイザ

スーパーバイザは3つの自動処理（漢字解読・単位分割・品詞認定）の結果をチェックし，その結果を修正するか，または処理のやり直しを命ずる。そこではそれぞれの処理によって得られた情報を利用する。すなわち，

1．文字種チェックの結果と漢字解読の結果はそれぞれの処理で利用する。

2．漢字解読で得た情報は単位分割に利用する。

すなわち，もし環境指定回路を適用したのなら，それは単語なので単位分割では分割しない。

3．漢字解読で得た情報は単位分割に利用する。

すなわち，もし仮名の小文字（ぁぃぅぇぉゃゅょっ）なら，プログラムはそこでは分割しない。

4．単位分割で得た情報は品詞認定で利用する。

すなわち，プログラムは単位分割において図4のテーブルを用いるが，そこで得られた品詞や活用情報は品詞認定に利用する。

処理結果のチェックは次の機能を含む。

1．助詞・助動詞の接続をチェックする。もしこれらの連続が日本語において不可能なら単位分割がミスをおかしたことになる。プログラムはこれらを修正する。

2．日本語においては助詞・助動詞を除いて1字で構成される単語は多くない。図8はいくつかの文字の頻度とその文字1字で構成する単語の頻度を示している。

助詞・助動詞でない高頻度語は単位分割の失敗によって作られたに違い

ない。プログラムはこのエラーを修正し，長い語に作り直す。

3．もし動詞の連用形が他の動詞に続いているのなら，それは複合動詞に違いない。プログラムはこれを修正し，長い語に作り直す。

図9はスーパーバイザの結果を示している。図9左のテスト文で，プログラムは，最初に図4のテーブルによって助動詞の連続として「た／ば／ね／ら／」と分割した。しかし，スーパーバイザはこの連続をチェックし修正する，また，品詞認定プログラムは図9のように動詞「たばねら」として情報をつけている。

文　字	頻　度	助詞助動詞の頻度	%	その他の語の頻度	%
の	38404	32588	84.9	2	0
い	23633	2	0.0	1305	5.5
し	22124	64	0.3	13138	59.4
に	18962	17037	89.8	3	0.0
と	16383	10173	62.1	0	0
は	16062	13324	83.0	0	0
た	15958	10569	66.2	1	0.0
る	15522	17	0.1	0	0
を	14710	14702	99.9	0	0
で	13515	8351	61.8	0	0

図8．文字の頻度とその1文字語の頻度

図9右のテスト文で，プログラムは「あそび／すぎ／た」と分割した。しかし，スーパーバイザはこれをチェックし，この語連続を複合語「あそびすぎ」に「た」が付いたものとして処理している。

①	②	③	④	⑤	⑥
沢山の木をたばねられませんでした。	たくさん	1	1	1	1
		1	R	1	R
	き	1	1	1	1
		1	R	1	R
		1	P+	0	
		1	R	0	
		1	Q	0	
				1	E8
		1	P#	1	P#
		1	P#	1	P#
		1	P+	1	P+
		1	P9	1	P9
		1	P+	1	P+
		1	Y	1	Y

①	②	③	④	⑤	⑥
面白くて遊び過ぎた。	おもしろ	1	M9E+	1	M9E+
		1	R	1	R
	あそ	1	E#	0	
	す	1	E#	1	E#
		1	P+	1	P+
		1	Y	Y	

図９．スーパーバイザの結果

(5) 活用形変換

文章中に現れた各活用形を終止形に変換する。同語異語の判別を助けるためのプログラムである。前後の文脈を調べないで終止形に変換するには，次の３つの方法がある。

処理方法	処理速度	辞書の大きさ	プログラム
① 活用語辞書とのマッチング	遅い	大きい	簡単
② 活用情報による終止形変換	早い	小さい	複雑
③ 出現形の漢字表記の利用	遅い	大きい	簡単

①は，活用語の語幹辞書を作り，そこに活用型と活用段の情報をつけ，これを利用する方法である。これには活用語辞書を検索する必要がある。

②は，このシステムで採用した方法だが，品詞認定の結果得られた活用情報により終止形に変換する方法である。ただし，未然形・連用形で自動変換できない語については辞書を作りこれを利用する。辞書検索が少ないだけ処理が早く，使用するメモリーも少なくて済む。

③は，入力データが漢字仮名混じりの場合に有効である。これは活用語の

漢字部分の最後の文字と仮名部分のローマ字表記から活用語尾を除いたものの辞書をつくる。たとえば，「動く・動か・動き・動け・動こ」は，辞書「動K」にまとめる。プログラムは，入力データから用意したローマ字語尾部分を削除し，辞書を引く。つぎに，辞書にある情報を付加する。この方法は①より辞書が小さくて済むという利点がある。しかし，入力データが漢字表記されていなければ正しく変換しない。

パソコンのようなメモリーの小さい，また処理速度の遅い機械では，②の方法が適当と考えてこれを採用した。

(6) 処理方法

入力データには，品詞認定の結果として活用形の情報が付いている。これを利用して次の処理を行う。

① 形容詞・助動詞は，プログラム内の活用表によって終止形に変換する。

② 動詞は，以下の方法による。

③ カ変・サ変は，プログラム内の活用表によって終止形に変換する。

④ 終止・連体形は，そのまま出力する。

⑤ 仮定・命令形は，語末の「れ・ろ・よ」を「る」に変える。それ以外は，語末をウ段に変える。

⑥ 未然形は，語末がエ段またはイ段なら「る」を加える。その他はウ段に変える。

⑦ 連用形は，語末がエ段なら「る」を加える。イ段または促音・撥音ならテーブルにしたがって変換する。たとえば，「いっ」はテーブルにしたがい，すべて「いく」と変換する。テーブルの内容は確率的に多い方を採用しておく。

(7) 語種認定

語種の認定は，漢字解読テーブルの読み情報を利用する。漢字解読テーブルの読み情報は，訓読みは平仮名，音読みは片仮名表記になっている。外来語読みはローマ字表記となっている。

漢字表記の語は，これらの情報を利用する。仮名表記の語は，片仮名なら

外来語，平仮名なら和語とする。

(8) 単語分かち書きデータの処理

単位分割を誤ると，品詞認定も活用形変換も誤ることになる。一貫処理の単位分割の精度は，後に述べる通り約90％である。品詞認定も活用形変換の精度は，この値にそれぞれの処理の精度を掛けた値となる。

しかし，はじめに述べた通り，それぞれの処理の結果は，KWICによって検査でき，その多くが一括して修正することが出来る。そこで，正しく分割されたデータを，品詞認定や活用形変換にかければ全体の精度はあがることになる。

このために，スペースで分割されたデータも処理できるようにしたのが，単語分かち書きデータの処理プログラム（NAPO）である。

このプログラムでは，単位分割の情報を受け取り，漢字解読・品詞認定・スーパーバイザ処理だけを行う。

(9) 文節分かち書きの仮名書きデータの処理

仮名書きされたデータも，文節分かち書きされていれば，語の認定処理が可能である。NAPKANAは，その機能を持ったプログラムである。文節末から付属語を切り出し，接続関係の判別に拠る品詞認定を行う。処理の制度は，現在７２％である。しかし，ＫＷＩＣによる修正で使用可能な水準となっている。

５．処理結果の精度

(1) はじめに

一貫処理プログラムの評価実験を大型計算機によって行った。パソコンのプログラムとほぼ同じだが，活用形変換・語種認定の機能はない。プログラムの作成は中野洋，その改良は石井正彦，処理および処理結果の検査は中野・石井・小沼悦が行った（文献３，７，30，31)。以下は，機械処理と人手作業との比較の結果についての報告である。報告の目的は，一貫処理を用いると，人手による処理作業と比べてどの点が良くなり，どの点が悪くなったか

をはっきりさせること，特に時間と精度について明らかにすることである。

図10は人手作業と機械処理の，それぞれの作業手順を「語彙調査データ作成の流れ」として，フローチャートで示したものである。

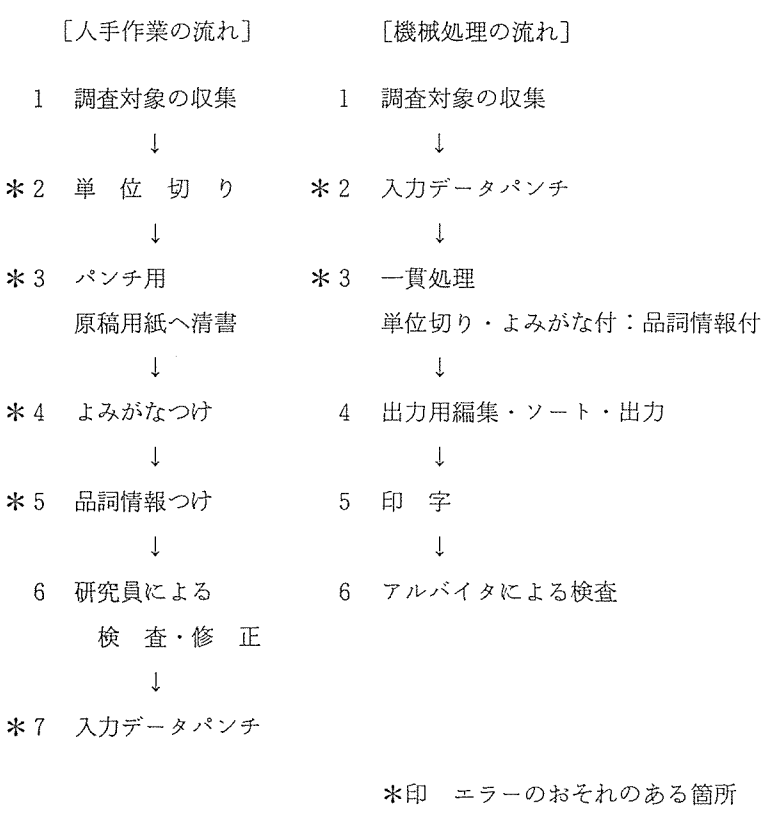

図10．語彙調査データ作成の流れ

(2) 調査対象について

調査対象をまとめたものが，次の表である。

分類	対象	総字数	漢字%	手作業者
高	世界史	2548	40.6%	大学2年生
教	政治経済	2067	37.2	――
校 科	物理	2353	30.8	――
書	生物	2642	33.3	――
雑	A．中央公論	5430	42.7	教育学部卒
	B．現代の眼	4787	31.5	――
誌	C．主婦と生活	4947	24.5	大卒(1データ)

高校教科書では，社会科と理科から2教科ずつ選択した。漢字含有率の異なる雑誌では，3冊からそれぞれ3データずつ9データをアトランダムに抽出している。字数は，高校教科書はそれぞれ2500字前後，雑誌は5000字程度をめやすとした。なお，漢字含有率の高低によって，自動分割の精度が影響されるのかどうかをみるために，その比率を算出した。この含有率は記号，スペース等を含んだ総数に対する割合である。最も含有率の高い<世界史>と，最も低い<雑誌C>とでは，約15%の差がある。

機械処理との比較実験のための人手作業の被験者は，通常この種の作業を手伝ってもらう大学生・大学卒業生の3名である。

(3) サンプル（世界史）

図11は　実際に人手作業をしてもらった<世界史>のサンプルである。教科書の原文にスラッシュの記号で単位切り作業をしている段階である。たとえば，「氏族村落から都市国家へ」という見出しを，作業者は「／氏族村落／から／都市国家／へ／」と単位切りをしているのがわかる。

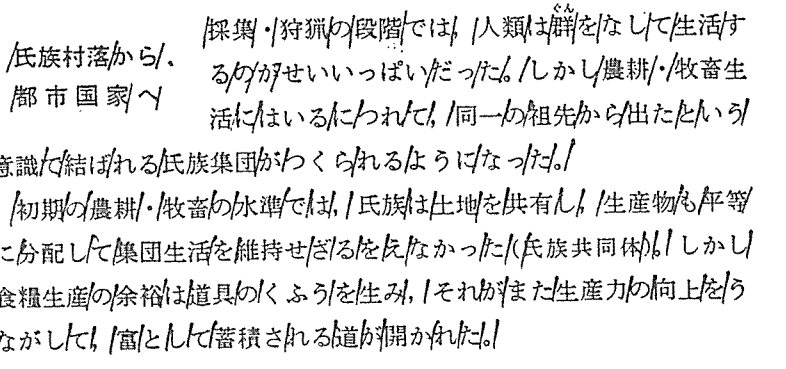

図11．サンプル（人手作業）

図12は，前と同じデータを機械処理した場合にどうなるか，その自動処理の結果である。図11の見出し，「氏族村落から都市国家へ」の部分を比較してみると，単位切りは正確になされているが，漢字解読では「都市国家」が「としこくか」となっていて，ミスをおかしているのがわかる。なお，はじめの数字は文の中の語末文字の位置を，最後の数字は文番号を示している。

004氏族村落	しぞくそんらく	名詞　00041
006から	から	助詞　00041
010都市国家	としこくか	名詞　00041
011へ	へ	助詞　00041
013採集	さいしゅう	名詞　00041
014・	・	記号　00041
016狩猟	しゅりょう	名詞　00041
017の	の	助詞　00041
019段階	だんかい	名詞　00041
020で	で	助助連用00041
021は	は	助詞　00041
022,	,	記号　00041
002人類	じんるい	名詞　00042
003は	は	名詞　00042
005群を	むれを	名詞　00042
006な	な	助助連体00042
007し	し	動詞連用00042
008て	て	助詞　00042
012生活する	せいかつする	動詞終連00042
014のが	のが	動詞未然00042
015せ	せ	助動未連00042
020いいっぱい	いいっぱい	名詞　00042
022だっ	だっ	助動連用00042
023た	た	助動終連00042
024.		記号　00042

図12．サンプル（自動処理結果）

(4) 精度について

機械処理に関して各精度を集計したものが，次の表である。

対象		機械処理		
		単位切り	漢字解読	品詞認定
教科書		90.6%	90.1%	96.9%
雑誌	A	93.1	89.0	96.7
	B	89.7	92.5	95.6
	C	88.0	87.2	95.0

高校教科書は4教科を一括して算出した。単位切り，漢字解読，品詞認定ともにほぼ90%以上の精度をあげている。雑誌は，漢字含有率からみると，A40%台，B30%台，C20%台のデータだが，その比率が高いほど単位切りの精度が良くなっている。品詞認定は，教科書・雑誌ともに95%以上の精度（ただし，この値は正しく単位切りされたものだけを対象とした）を示している。

(5) 人手と機械の精度について

人手作業と機械処理の精度を比較したものが，次の表である。

	世界史			雑誌	
	機械	修正後	人手	機械	人手
単位切り	92.7%	97.0%	97.0%	91.4%	99.2%
よみがな	92.2	99.9	100.0	87.9	99.7
品詞認定	97.3	93.5	91.7	96.0	99.9

機械処理および人手作業でも，約90%以上の精度を見込むことができる。＜世界史＞のデータでは，機械処理をしたあと同じ作業者に少し期間をおいてから修正作業をしてもらった。その結果，単位切り・よみがなつけは修正によって精度がアップしたが，品詞認定はダウンしている。これは，機械処理では，品詞の認定基準が精度を計算する上で少し甘くなっているためで，

たとえば「で」には，格助詞の「で」と助動詞の「で」とがあるが，機械処理での精度の計算では，どちらかの情報が与えられている場合はエラーとみなさないことにした。ところが，人手作業による修正では正確な情報以外は認めないようにしたのである。

(6) 処理時間について

作業における処理時間をまとめたものが，次の表である。

	機械処理			人手作業					
	一貫処理	全体*	検査	単位	清書	かな	品詞	全体**	検査
世界史	0.1秒	30分	4時間	2時間	5時間	2時間	6時間	3日	1時間
雑誌4ータ	0.6秒	64分	———	2時間	11間	7時間	9時	6日	1時間

* LOGIN，LOGOUT，オペレートミスなどのすべてを含む

** 仕事につく前の時間・休憩なども，すべて含む

機械の処理時間では，オペレートを開始した時間からデータを印字し終わるまでの全体の時間を計算してある。オペレートの慣れ，不慣れによっても所要時間が多少違ってくる。＜世界史＞の場合，全体の処理時間は３０分，またこの機械処理で出たデータをアルバイタによって修正作業をした時の所要時間は４時間となった。よって，機械と人手をあわせた全部の作業としては４時間３０分かかっているわけである。ちなみに一貫処理だけの所要時間というのは世界史で0.1秒，雑誌で0.6秒であり，あっという間に終わってしまう。これが人手作業だけとなると，全体の作業日数としてほぼ３日かかる。なお，そのあとの研究員による修正検査時間は１時間であった。同じように＜雑誌＞のほうは，機械処理では全体で６４分のところを，人手作業だけでは，全体の作業が終了するまで６日もかかっているということがわかる。

ただし，機械処理については，データの入力に要する時間，修正に要する

時間を考えに入れなければならない。前者については，光学文字読取り装置の利用，電算写植用データの利用，外部依託も考えられる。後者については，効率的な修正システムを開発しているが，別に機会を得て報告したい。しかし，機械でも人手でも図10に示した通り，入力・修正は必要である。

(7) パンチ量について

パンチ量を比較したものが，次の表である。

	語数	機　械	人　手	割合　機／人
世界史	1296	2548字	11926字	21.4%
雑誌A	4217	7596字	37531字	20.2%

＜世界史＞の語数は1296語である。文字数に直すと2548字である。機械処理では入力データはこれだけでよい。ところが，人手作業では原データによみがなや品詞情報をつけるので，11926字になる。その割合は機械は人手の約21.4%になっている。＜雑誌A＞についても同じようなことが言える。

(8) まとめ

1．語彙調査データの作成作業における人手の作業と機械処理の比較を行った。

2．処理精度は，単位切りでは機械でほぼ90%，人手では97%～98%が見込まれることがわかった。これは明らかに人手の方がよい。

3．処理時間は検査の時間を含めても機械が約５時間，人手が約53時間であり，人手は機械処理の10倍以上かかっているのがかる。

4．入力パンチ量については，機械は人手の約２０%の入力で済む。

5．以上の結果として今後の語彙調査には機械による自動処理を用いても良いことは明らかである。しかし，今まで以上によい修正システムをつくる必要があると思われる。

6．プログラム一覧

一貫処理システムは，次表のプログラムによって構成されている。

表中の①②③が一貫処理の本体である。プログラムは，①で入力データを読み込み漢字解読処理を行う。次に，その処理結果を②で，単位切り，語形による品詞認定を行い，スーパーバイザでチェックし，単位切り情報を修正する。これは，多くの場合単位切りが短く切りすぎるので語を長くする方向での修正である。チェックが通らなければ，単位切り・品詞認定を１０回まで繰り返す。１０回のうちにチェックが通らなければ，そのままを③に渡す。

	プログラム名	内　　容
①	KAIDOK.EXE	漢字解読
②	SUPER.EXE	単位切り，品詞認定１，スーパーバイザ
③	PARTS2.EXE	品詞認定２
④	OUTPUT.EXE	清書出力
⑤	NAP.EXE	一貫処理ドライバ
⑥	NAPOUT.EXE	一貫処理ドライバ清書出力付
⑦	KAIDOKO.EXE	漢字解読スペース分かち書き用
⑧	PARTS.EXE	品詞認定スペース分かち書き用
⑨	NAPO.EXE	スペース分かち書き用一貫処理ドライバ
⑩	NAPKANA.EXE	仮名文節分かち書き用
⑪	CONJ.EXE	活用形変換
⑫	GOSYU.EXE	語種認定

②の処理結果を用い，③で接続による品詞認定を行う。ここでは，品詞情報のチェックと修正を行う。

④の清書出力は，一貫処理の結果を単語単位で出力し，品詞や活用コードを漢字で表示するプログラムである。これまでの図示の通り，文字レコードとしてのものであるため，処理結果がみにくいので，みやすくしたものである。

一貫処理システムは，以上の通り，①②③⑪⑫とプログラムが分れている。したがって，入力フォーマットが一致するかぎり，それぞれ独立して用いることが出来る。⑤⑥は，①②③および④を連続して用いるためのものである。

⑦⑧⑨は，スペース分かち書きした漢字仮名混じり文（一貫処理の結果をＫＷＩＣファイルで修正し，原文作成プログラムにかけると，こうなる。単位切りの結果が全体の処理の精度を左右するためにこのプログラムを作った。）を処理するプログラムである。単位切り処理は行なわない。

⑩は，文節分かち書きの仮名データを処理するためのプログラムである。

⑪は活用形変換，⑫は語種認定のプログラムである。

以上を図示すれば，次の通りである。

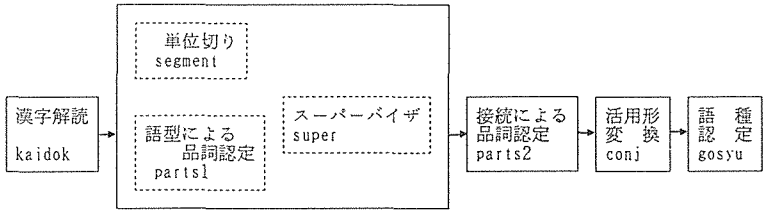

図13　システム構成

動作環境

MS-DOSが動き漢字の使えるコンピュータであれば使用できるはずである。メモリーは空き領域128K以上あればよい。補助記憶装置は，フロッピーディスク１台でも動作可能である。しかし，出力ファイルが入力ファイルの８倍の大きさになり，又，中間ファイルも出力ファイルと同じだけの領域を必要とする。大量データを処理する場合は，フロッピーディスク２台，できれば固定ディスク装置があったほうがよい。

一貫処理システムのそれぞれのプログラムの入出力は基本的には標準入力（キーボード）から入力し，標準出力（画面）に出力するようになっている。この時の入力の終わりは，EOF(CTRL/Z) である。入力データは「,」または

「。」で区切られた最大２５６文字の文字データである。しかし，起動時のパラメータによって入出力ファイルを指定することが出来る。方法は全プログラムに共通で次の通りである。

A＞program [inputfile] [outputfile]

program: それぞれのプログラム名

inputfile: 入力ファイル名 省略できる．省略時は標準入力

outputfile: 出力ファイル名 省略できる．省略時は標準出力

一般的な使い方は，次の通りである。ただし，以下の｜の記号は16進数字で７Ｃにあたる文字である。

A＞KAIDOK inputfile ｜ SUPER ｜ PARTS2 ＞ outputfile

上記の例ではinputfileを入力としてoutputfileに出力する．

又，NAP.EXEというプログラムを使えば次の命令だけで良い。

A＞NAP inputfile outputfile

NAPKANA.EXEの使い方。

NAPKANA.EXEは，従来のKAIDOK.EXEとSUPER.EXEが一緒になったようなもので，

A＞KAIDOK inputfile ｜ SUPER ｜ PARTS2 ｜ OUTPUT ＞ outputfile

を実行する。使い方としては，次の通り。

A＞NAPKANA inputfile ｜ PARTS2 ＞ outputfile

又は

A＞NAPKANA inputfile ｜ PARTS2 ｜ OUTPUT ＞ outputfile

７．辞書一覧

(1) 辞書一覧

次頁の表に，一貫処理で使用した主要な辞書の一覧を示す。これらは，MS-DOSのテキストファイルで書かれている。他の日本語処理の辞書に比べ，大変小さいことが特徴となっている。また，辞書は，処理対象の文章に応じて

書き換えることができ，処理の精度を上げることが出来る。

	辞　書　名	ファイル名	バイト数	項目数
①	漢字解読用テーブル	KAN.TBL	94477	2945
②	漢字解読テーブル用索引	KAN.IDX	17672	
③	単位切りテーブル（漢字仮名混じり用）	SEGMENT.TBL	10321	359
④	単位切りテーブル（仮名分かち書き用）	SEGREV.TBL	10327	359
⑤	助詞，助動詞接続テーブル	POSTBL1.TBL	1357	34
⑥	品詞接続テーブル	POSTBL2.TBL	297	15
⑦	助詞，助動詞接続チェック用テーブル	PRTSTR.TBL	6391	142
⑧	連用形変換テーブル	RENYOU.TBL	103797	3660

(2)　辞書の変更

一貫処理プログラムでは，プログラム部分と辞書部分を出来るだけ独立させている。とくに，辞書には処理の基準が書き込まれている。したがって，辞書の内容を変更することで処理基準を変えることができる。とくに，単位切りテーブルは単位切りの精度を左右するものであり，かつこれは入力データによって変更したほうがよい性格のものである。上記のテーブルは，新聞の語彙調査データを用いて作成したものだから比較的仮名連続や助詞助動詞連続が少ないとおもわれる。同じ著者による文章であれば，文字遣いや口調などをテーブル化することによって精度をあげることができると思われる。辞書を変更した場合，指定の順序にソートしなおさなければならない。

(3)　辞書の内容の例

仮名ローマ字変換テーブル(KANARO.TBL,1015バイト)

ぁ＊A	う＃U	かKA
あ＃A	ぇ＊E	がGA
ぃ＊I	え＃E	きKI
い＃I	ぉ＊O	ぎGI
ぅ＊U	お＃O	くKU

漢字解読用テーブル(KAN.TBL,94477バイト)

```
1亜ア　　☆
1啞ア　　☆
1娃アイ　☆
1阿ア　　☆
3哀１１アイ　　２Ａあわ　　３８かな　　＊んれ２んし３☆
1愛アイ　☆
1挨アイ　☆
1姶ア　　☆
1逢あ　　☆
1葵あおい　☆
1茜あかね　☆
1穐あき　☆
3悪１１あく　　２８お　　３Ａわる　　４８あ
＊Ｎど１Ｍ嫌２Ｍ憎２Ｎい３Ｎく３Ｎ者３☆
2握１アク　　Ａにぎ　　☆
1渥あつい　☆
2旭Ａあさひ　１キョク　☆
1葦あし　☆
1芦あし　☆
```

単位切りテーブル（漢字仮名混じり用）（SEGMENT.TBL,10321バイト）

```
3および　　3Ａ
3おける　　2ＥＲ1ＰＩ
4こうした　　2Ｃ　1Ｅ９1Ｐ＋
2して　　1Ｅ９1Ｒ
1た　　1Ｐ＋
```

品詞と活用の部分は，コードで表わしていて，その意味は次のようになっている。

品詞コード

1　名詞

A　接続詞　　B　感動詞

C　副詞　　D　連体詞

E　動詞　　M　形容詞

P　助動詞　　Q　助動詞または助詞　　R　助詞

Y　記号　　X　数字

活用コード

8　未然形　　9　連用形　　#　未然形または連用形

H　終止形　　I　連体形　　+　終止形または連体形

Q　仮定形　　R　命令形

助詞，助動詞接続テーブル(POSTBL1.TBL,1357)

```
の@WR1　@# と# から# で# へ# より# まで# だけ# ばかり# こそ# など# ぐらい#
 I +1 /1 @ @
を@WR1　@# と# から# まで# の# だけ# ばかり# こそ# さえ# すら# のみ# など#
ぐらい# 1 /1 @ @
に@WR1　@# と# の# だけ# ばかり# のみ# など# ぐらい# 494# I +1 /1 @ @
は@WR3　@# と# に# から# で# より# まで# の# だけ# ばかり# こそ# など# ぐら
い# 9 # + I1 /1 @ @
が@WR13　@# の# と# から# まで# も# だけ# ばかり# こそ# さえ# のみ# など# ぐ
らい# I + H1 /1 @ @
と@WR13　@# の# だけ# ばかり# のみ# など# ぐらい# H I + R1 /1 @ @
で@WR13　@# ない# 91 /1 @ @
た@WP 1+ @ E9 9/E9@ @
```

品詞接続テーブル(POSTBL2.TBL,297バイト)

```
@ Y    @#か# さ# ぞ# ね# よ#  H +@ @
@ T    @#が# て# し#   9 #  I   +A B @ @
@ 1    @ M 4A  1 @ @
@ E    @#て# ては# ても#   M9M#E9E1@ @
```

このデータのフォーマットは次の通りである。

```
1   2      ┌──────────────── 3 ──────────────────────────
の@WR1   @# と# から# で# へ# より# まで# だけ# ばかり# こそ# など#
────┐ 4      5
ぐらい#  I  +1  /1 @ @
```

1．語

2．品詞

3．この語の直前に用いることのできる助詞・助動詞

4．この語の直前に用いることのできる品詞と活用形

5．もし，直前の語が３，４と一致しなければ強制的に適用する品詞・活用形

助詞，助動詞接続チェック用テーブル(PRTSTR.TBL,6391ﾊﾞｲﾄ)

らしい	` 1                  1      2           111`
そうだ	`                               1 1     121`
う	`  1                        2   1 1 1   221`
よう	`                   11      2        1 121`
る	`              1                        2`
ようだ	`                           2            2`
ようです	`                           2       1    2`
れれ	
なけれ	
べき	`               2211                 1  1     2`

縦に並んだ付属語と横に並んだ付属語（図では明示していない）との交差する箇所に1 または2 があればその接続が存在することを示す。2は頻度が多いことを示しているがプログラムではこの情報を使用していない。

連用形変換テーブル（RENYOU.TBL,103797,3660）

五段あい	う	五段あおのき	く
五段あい	く	五段あおみ	む
五段あいつい	ぐ	五段あおむい	く
五段あいつぎ	ぐ	五段あおむき	く
五段あえい	ぐ	五段あおり	る
五段あえぎ	ぐ	五段あおん	む
五段あおい	ぐ	上一あかさび	る
五段あおぎ	ぐ	五段あかし	す
五段あおっ	る	五段あかしくらし	す
五段あおのい	く	五段あからみ	む

その語形があれば，最後の文字を右端の一字に変える。

8．今後の課題

機械の発達によりシステムは変化すると考えられる。しかし，言語処理部分は変らない。したがって，言語処理部分を独立させ，新しい機械システムに組み込む努力と注意が必要である。また，完全な機械処理が望めないとすると，いかに効率的な修正システムを作るかが重要となる。一般に機械処理の誤りの箇所は特定できない。しかし，経験的に誤りの起こりやすい箇所は予想できる。そのような箇所を抽出し，注意して検査するためのシステム，作業を繰り返しやすくするためのファイル管理システムなどが重要となろう。

今後ますます高速大容量の機械が出現するに違いない。しかし，上に述べたようにわれわれの研究は，人間の検査なしには考えられないのである。ヒューマンインターフェースに優れた機械の出現を望みたい。

本プログラムシステムおよび辞書をフロッピーでご入用の方は筆者まで文書でご連絡ください。

謝　辞

学習院大学の田中章夫教授は国語研究所に在職されていた時に第4研究部第1資料研究室長として，この研究の計画を立案され，かつ，漢字解読のプロトタイプを作成された。江川清情報資料研究部長は，第4研究部第1資料研究室におられた時に単位分割のプロトタイプを作成された。いずれも，本プログラムにそれらを利用することを許された。

また，一貫処理プログラムの評価については言語計量研究部第1研究室（当時）の石井正彦研究員と小沼悦研究補助員の，パソコンへの移植についてはアルバイタ山田雅一氏の協力を得た。

以上の方々の他，香川大学土屋信一教授，上越教育大学鶴岡昭夫助教授，国立国語研究所山崎誠研究員，筑波大学の荻野綱男講師の貴重な意見を得ることができた。それぞれ記して，感謝の意を表する。

参考文献

1．荒木　啓介・板山　和彦「JICSTの実用的漢字―カナ変換システムK-KACSについて」（情報処理20－10，1979）

2．石井　久雄「「一貫処理システム」KAIDOKU プログラムの『雑誌用語の変遷』語彙表への適用」（ＣＬ通信第3号，1989.3，国立国語研究所言語　計量研究部）

3．石井　正彦「自動単位分割の精度と問題点」（ＣＬ通信第3号，1986.4，10，国立国語研究所言語計量研究部）

4．岩波書店編「電子広辞苑」（岩波書店，1988）

5．江川　清「漢字かな混り文の「自動単位分割」に関する一研究」（計量国語学第43／44巻，1968）

6．江川　清「単位分割自動化のシステムについて」（計量国語学第51巻，1969）

7．小沼　悦「一貫処理プログラムの評価実験(3)――精度，人手作業との比較において――」(CL通信第3号，1986.4，国立国語研究所言語計量研究部)
8．国語学会・国立国語研究所編「日本語研究文献目録・雑誌編」[フロッピー版](秀英出版，1989)
9．国立国語研究所「電子計算機による新聞の語彙調査」(国立国語研究所報告37，秀英出版，1970)
10．国立国語研究所「高校教科書の語彙調査」(国立国語研究所報告76，秀英出版，1983)
11．国立国語研究所「中学校教科書の語彙調査」(国立国語研究所報告87，秀英出版，1986)
12．斎藤　秀紀「漢字かな混り文の文字列」(LDP月報別冊8，1971,2，国立国語研究所)
13．坂本　義行「文節単位の自動分割法－字種と平仮名連糸による－」(計量国語学11巻6号，1978,9)
14．真田　治子「文体の自動変換－ダ体からデス・マス体へ」(計量国語学16巻7号，1988，12)
15．三省堂編「模範六法CD-ROM版」(三省堂，1989)
16．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，秀英出版)
17．藤崎博也・亀田弘之「自動単位切りによる新聞記事の語彙調査」(昭和60年度文部省科学研究費特定研究(1)「情報化社会における言語の標準化」報告書，1986)
18．宮島達夫・中野洋・鈴木泰・石井久雄「フロッピー版古典対照語い表および使用法」(笠間書院，1989)
19．バイブルズ編「CD-HIASK（朝日新聞前文データベース)」(CD-ROM電子出版サービス「バイブルズ」，1989)
20．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，秀

英出版）
21. ――「索引作成プログラムライブラリ」（電子計算機による国語研究Ⅶ，1976，秀英出版）
22. ――「索引作成プログラムライブラリ」（電子計算機による国語研究Ⅷ，1977，秀英出版）
23. ――「言語研究における一貫処理の研究」（電子計算機による国語研究Ⅹ，1978，秀英出版）
24. NAKANO Hiroshi, TSUTIYA Shin'iti, TURUOKA Akio 「AN AUTOMATIC PROCESSING OF THE NATURAL LANGUAGE IN THE WORD COUNT SYSTEM 」(Proceedings of The 8th Internatonal Conference on Computational Linguistics,1980)
25. ――「分類番号つけ支援システム」（情報処理学会計算言語研究会資料，1981,2）
26. ――「話しことばの語彙調査」（情報処理学会自然言語処理研究会資料，1982,3）
27. ――「ひらがなの使用頻度とひらがな一字で表記される語の頻度数」（季報1980－夏号，国立国語研究所言語計量研究部）
28. ――「言語研究のためのデータベース　データベース操作」（昭和55年～57年度文部省科学研究費一般研究（A）「話しことばの計量国語学的調査・分析のための基礎的研究」研究報告書（第２分冊），1983,3）
29. ――「語彙調査の自動化における一貫処理システム」（CL通信第１号，1985，国立国語研究所言語計量研究部）
30. ――「自動漢字解読の精度と問題点」（CL通信第３号，1986.4，国立国語研究所言語計量研究部）
31. ――「自動品詞認定の精度と問題点」（CL通信第３号，1986.4，国立国語研究所言語計量研究部）

付録１：漢字解読テーブル(KAN.TBL,94477ﾊﾞｲﾄ)

```
1〇レイ　　☆
1亜ア　　☆
1啞ア　　☆
1娃アイ　　☆
1阿ア　　☆
3哀１１アイ　　２Ａあわ　　３８かな　　＊Ｎれ２Ｎし３☆
1愛アイ　　☆
1挨アイ　　☆
1姶ア　　☆
1逢ア　　☆
1葵あおい　☆
1茜あかね　☆
1穐あき　　☆
3悪１１アク　　２８オ　　３Ａわる　　４８あ　　＊Ｎど１Ｎい３Ｎく３Ｎ者
3☆
2握１アク　　Ａにぎ　　☆
1渥アク　　☆
2旭Ａあさひ　１キョク　☆
1葦あし　　☆
1芦あし　　☆
1鯵あじ　　☆
2梓Ａあずさ　１シ　　☆
2圧１アツ　　Ａお　　☆
1斡アツ　　☆
1扱あつか　☆
1宛あて　　☆
1姐あね　　☆
1虻あぶ　　☆
1飴あめ　　☆
2絢Ａあや　　１ケン　　☆
1綾あや　　☆
1鮎あゆ　　☆
1或あ　　☆
2粟Ａあわ　　１ゾク　　☆
1袷あわせ　☆
3安１１アン　　２Ａやす　　＊Ｎ物２Ｎ売２Ｎい２Ｎく２☆
2庵１アン　　Ａいおり　☆
1按アン　　☆
3暗１３アン　　２Ｃくら　　＊Ｍ薄２Ｍ真２Ｎい２Ｎら２☆
1案アン　　☆
1闇やみ　　☆
2鞍１アン　　Ａくら　　☆
1杏キョウ　☆
2以１イ　　Ａもつ　　☆
1伊イ　　☆
3位１１イ　　２Ａくらい　＊ＭＳ１Ｍ気２☆
2依１イ　　Ａよ　　☆
2偉１イ　　Ａえら　　☆
3囲１１イ　　２Ａかこ　　＊Ｎい２☆
2夷１イ　　Ａえびす　☆
```

3委１１イ　　２８ゆだ　　３Ａまか　　＊Ｎか３Ｎせ３Ｎね３☆
１威イ　　☆
１尉イ　　☆
１意イ　　☆
３慰１１イ　　２Ａなぐさ　＊Ｎみ２Ｎめ２☆
３易１６イ　　２３エキ　　３Ａやす　　＊Ｍ簡１Ｎ貿２Ｎい３Ｎく３☆
１椅イ　　☆
２為１イ　　Ａため　　☆
２畏１イ　　Ａおそ　　☆
２異１イ　　Ａこと　　☆
３移１１イ　　２Ａうつ　　＊Ｎり２Ｎる２Ｎっ２Ｎら２☆
１維イ　　☆
１緯イ　　☆
１胃イ　　☆
２萎１イ　　Ａしお　　☆
３衣１１イ　　２８エ　　３Ａころも　＊Ｍ単２Ｍ羽３Ｍ紋２☆
２謂１イ　　Ａイ　　☆
３違１３イ　　２Ｃちが　　３８たが　　＊Ｍ相１Ｍ差１Ｍ仲３☆
３遺１１イ　　２８ユイ　　３Ａのこ　　＊Ｎ言２☆
１医イ　　☆
３井１８セイ　　２Ｈい　　３８ジョウ　＊Ｍ天３Ｍ市１Ｍ油１Ｎ田１☆
１亥イ　　☆
１域イキ　　☆
３育１１イク　　２Ａそだ　　３８はぐく　＊Ｎち２Ｎて２Ｎむ３Ｎん３☆
１郁いく　　☆
１磯いそ　　☆
３一１Ｈイチ　　２８イツ　　３８ひと　　４８イッ　　＊Ｍ統２Ｍ同２Ｍ唯２Ｎつ
３Ｎ般４☆
１壱イチ　　☆
２溢１イツ　　Ａあふ　　☆
２逸１イッ　　Ａはや　　☆
３稲１８トウ　　２７いね　　３Ｇいな　　＊Ｍ水１Ｍ陸１Ｎ苗１☆
１茨いばら　☆
１芋いも　　☆
１鰯いわし　☆
１允イン　　☆
３印１１イン　　２Ａしる　　＊Ｍ旗２Ｍ目２Ｍ矢２Ｎし２☆
１咽イン　　☆
１員イン　　☆
３因１Ｈイン　　２８ちな　　３８よ　　＊Ｎみ２Ｎり３Ｎっ３Ｎる３☆
１姻イン　　☆
３引１１イン　　２Ａひ　　３８ひき　　＊Ｍ取３Ｎき２Ｎ上３Ｎ下３☆
３飲１１イン　　２Ａの　　＊Ｍ酒２Ｎみ２☆
１淫イン　　☆
１胤イン　　☆
１蔭かげ　　☆
１院イン　　☆
２陰１イン　　Ａかげ　　☆
３隠１１イン　　２８オン　　３Ａかく　　＊Ｎ密２Ｎ亡２Ｎれ２☆
１韻イン　　☆

（以下略）

付録２：単位切りテーブル（漢字仮名混じり用）

（segment.tbl,10321ﾊﾞｲﾄ）

５．ところが		１Ｙ　４Ａ	１が	１Ｒ	
５．ところで	１Ｙ	４Ａ	４きちんと	４Ｃ	
３．だが	１Ｙ	２Ａ	３きっと	３Ｃ	
３．所で	１Ｙ	２Ａ	３きょう	３１	
３．所が	１Ｙ	２Ａ	４ください	４Ｐ	
３．でも	１Ｙ	２Ａ	３くらい	３Ｒ	
３．では	１Ｙ	２Ａ	３くらい	３Ｒ	
４あくまで	４Ｃ		４けれども	４Ｒ	
４あらたに	４Ｃ		３けれど	３Ｒ	
４あまりに	４Ｃ		４こうした	２Ｃ	１Ｅ９１Ｐ＋
３あなた	３１		４こまかに	４Ｃ	
３あまり	３Ｃ		３ことば	３１	
２あと	２１		３これら	３９	
２あれ	２１		３こども	３１	
２あす	２１		３ことし	３１	
２あの	２Ｄ		３ことに	３Ｃ	
５いちがいに	５Ｃ		３こんな	３Ｄ	
５いっきょに	５Ｃ		３こんど	３１	
５いやしくも	５Ｃ		２こう	２Ｃ	
５いっぱんに	５Ｃ		２こと	２１	
５いたずらに	５Ｃ		２この	２Ｄ	
５いっこうに	５Ｃ		２ここ	２１	
５いちように	５Ｃ		２こそ	２Ｒ	
４いってに	４Ｃ		２これ	２９	
４いわんや	４Ｃ		２この	２Ｄ	
４いっきに	４Ｃ		２ごろ	２１	
４いかなる	４Ｄ		６さっきゅうに	６Ｃ	
３いかに	３Ｃ		３させる	３Ｐ＋	
３いたし	３Ｅ		３させよ	３ＰＲ	
３いずれ	３Ｃ		３さらに	３Ｃ	
３いかん	３Ｃ		３させれ	３ＰＱ	
３いかが	３Ｃ		３さらに	３Ｃ	
２いう	２Ｅ＋		２させ	２Ｐ＃	
２いる	２Ｅ＋		２さえ	２Ｒ	
２うえ	２１		２さる	２Ｄ	
２うち	２１		５したがって	４Ａ	
５おのずから	５Ｃ		４しばしば	４Ｃ	
５おおいなる	５Ｄ		４しきりに	４Ｃ	
４おのずと	４Ｃ		４しだいに	４Ｃ	
４おおいに	４Ｃ		３しめよ	３Ｐ＋	
３および	３Ａ		３しかし	３Ａ	
３おける	２ＥＲ１ＰＩ		３しかも	３Ａ	
２おり	２Ｅ９		３しめる	３Ｅ＋	
２おい	２Ｂ		３しかも	３Ａ	
６かならずしも	６Ｃ		３しめろ	３ＰＲ	
３かなり	３３		２して	１Ｅ９１Ｒ	
３かりに	３Ｃ		２しめ	２Ｐ＃	
２から	２Ｒ		２しか	２Ｒ	
１か	１Ｒ		５じょじょに	５Ｃ	

3じつに	3C
6すくなくとも	6C
4すなわち	4C
3すでに	3C
3すべて	3C
3すでに	3C
3ずっと	3C
1ず	1P9
2せる	2P＋
2せれ	2PQ
2せろ	2PR
2せよ	2PR
1せ	1P＃
5ぜったいに	5C
6そうきゅうに	6C
4そうだろ	4P8
4そうなら	4PQ
4そうごに	4C
4そうだっ	4P9
4それぞれ	41
4それとも	4A
3そんな	3D
3そうな	3PI
3そうで	3P9
3そして	3A
3そうだ	3PH
3そうに	3P9
2その	2D
2そう	2C
2そこ	21
2それ	29
3ぞくに	3C
4たんなる	4D
4ただちに	4C
4たがいに	4C
3たかろ	3P8
3たけれ	3PQ
3たかっ	3P9
3たんに	3C
3ただに	3C
2たら	2PQ
2たい	2P＋
2たく	2P9
2ただ	2C
2たり	2R
2ため	21
2たろ	2P8
1た	1P＋
3だから	3A
2だっ	2P9
2だけ	2R
2だろ	2P8
2だり	2R
1だ	1PH
3つまり	3C
3ついに	3C
2つい	2C
2つぎ	21
4ていない	1R　1E＃2P
2ても	2R
1て	1R
3でしょ	3P8
2でし	2P9
2でき	2E＃
2でる	2E＋
2です	2PH
2でも	2R
1で	1Q9
6とんでもない	6M
4とにかく	4C
4ともかく	4C
4ところが	4A
3とくに	3C
3ともに	3C
3ところ	31
2とも	2R
2とき	21
2とう	2P9
1と	1R
2どこ	21
2どう	2C
2どの	2D
4ならびに	4A
3なかっ	3P9
3ながら	3R
3なかろ	3P8
3なけれ	3PQ
2なあ	2R
2なか	21
2など	2R
2なく	2P9
2なり	2R
2なお	2C
2なら	2PQ
2ない	2PH
2なる	2E
2なぜ	2C
1な	1Q1
4に対する	1R　3E＋
4にわかに	4C
1に	1R
1ぬ	1P＋
1ね	1QQ
2ので	2R
2のち	21
2のに	2R

1の	1R
4はるかに	4C
4はないか	1R　2E＋1R
4はじめて	4C
2はず	21
1は	1R
3ばかり	3R
1ば	1R
1ば	1R
5ひじょうに	5C
4ひいては	4C
4ひそかに	4C
4ひとえに	4C
3ひとつ	31
3ふいに	3C
3ふたつ	31
1へ	1R
4ほとんど	4C
2ほか	21
2ほど	2R
4まことに	44
4まもなく	4C
3まさに	3C
3または	3A
3ましょ	3P8
3ますれ	3PQ
2まだ	2C
2また	2A
2まで	2R
2ませ	2P井
2まい	2P＋
2まま	21
2まず	2C
2ませ	2P8
2まし	2P9
2ます	2P＋
4みだりに	4C
3みんな	31
4もっとも	4C
4もちろん	4C
3もっと	3C
3もっと	3C
2もの	21
2もと	21
2もう	2C
1も	1R
3やはり	3C
3やっと	3C
3やがて	3C
2やら	2R
1や	1R
4ようだろ	4P8
4ようだっ	4P9
4ようなら	4PQ
4ようやく	4C
3よほど	3C
3ようだ	3PH
3ような	3PI
3ようで	3P9
3ように	3P9
2よう	21＋
2より	2R
2よる	2E＋
4らしゅう	4P9
4らしかっ	4P9
4らしけれ	4PQ
3らしく	3P9
3られよ	3PR
3らしい	3P＋
3られれ	3PQ
3られる	3P＋
2られ	2P井
2られ	2P井
2れよ	2PR
2れれ	2PQ
2れろ	2PR
2れる	2P＋
1れ	1P井
4われわれ	41
3わりに	3C
3わざと	3C
2わが	2D
1を	1R
1ん	1P井
2或は	2A
3一気に	3C
3一挙に	3C
3一様に	3C
3一手に	3C
3一般に	3C
2一段	2C
2一層	2C
2一体	2C
2仮に	2C
1何	1C
2我が	2D
3確かに	34
2割に	2C
4間もなく	4C
2及び	2A
2共に	2C
3決して	3C
2結構	2C
2結局	2C
2現に	2C
3互いに	3C

3交互に	3C
2再び	2C
2最も	2C
3細かに	3C
2時々	2C
3次第に	3C
3次いで	3A
2次に	2A
2自ら	2C
2実は	2C
2実に	2C
2主な	2D
2殊に	2C
3従って	3A
3初めて	3C
5少なくとも	5C
2少々	2C
1尚	1C
3新たに	3C
3新しい	3M
2真に	2C
3正しい	3M
2正に	2C
3絶対に	3C
2全く	2C
3早急に	3C
3相互に	3C
2即ち	2A
2俗に	2C
4大いなる	4D
3大きな	3D
3大いに	3C
2但し	2A
3単なる	3D
3単なる	3D
2単に	2C
2単に	2C
3直ちに	3C
3同じく	3C
2同じ	2C
2特に	2C
2特別	2C
3非常に	3C
2必ず	2C
3不意に	3C
3並びに	3A
2又は	2A
1又	1A
3例えば	3C

付録３：助詞・助動詞接続テーブル（POSTBL1.TBL,1357ﾊﾞｲﾄ）

```
の@WR1　@# と# から# で# へ# より# まで# だけ# ばかり# こそ# など#
ぐらい#　I +1 /1 @ @
を@WR1　@# と# から# まで# の# だけ# ばかり# こそ# さえ# すら# のみ
# など# ぐらい# 1 /1 @ @
に@WR1　@# と# の# だけ# ばかり# のみ# など# ぐらい# 494# I +1 /1
@ @
は@WR3　@# と# に# から# で# より# まで# の# だけ# ばかり# こそ# な
ど# ぐらい#　9 # + I1 /1 @ @
が@WR13　@# の# と# から# まで# も# だけ# ばかり# こそ# さえ# のみ#
など# ぐらい#　I + H1 /1 @ @
と@WR13　@# の# だけ# ばかり# のみ# など# ぐらい#　H I + R1 /1 @ @
で@WR13　@# ない#　91 /1 @ @
た@WP 1+ @ E9 9/E9@ @
も@WR3　@# に# を# と# から# で# へ# より# まで# の# さえ# すら# な
ど# ぐらい#　9 # I +1 /1 @ @
から@WR13　@# の# だけ# ばかり# など# ぐらい# のみ# 1　H/1 @ @
な@WP 2I @# の# 1 /1 @ @
れ@WP　# @ E8/E8@ @
ない@WP 3H @# せ# させ# れ# られ# だがら# は# E8/E @ @
て@WR2　@　9/E9@ @
だ@WP 2H @# の# 1 /1 @ @
まで@WR3　@# の# など# ぐらい#　I +1 @ @
ます@WP 4+ @# せ# させ# れ# られ# たがり# E9/E9@ @
へ@WR1　@ 1 /1 @ @
です@WP 4H @ M11 /1 @ @
や@WR1　@ 1 /1 @ @
れる@WP　+ @ E8/E8@ @
う@WP　+ @# たろ# でしょ# だろ# E8M848/E8@ @
られ@WP　# @ E8/E8@ @
ように@WP 59 @# た#　I/EI@ @
だっ@WP 29 @# の# 1 /1 @ @
なく@WP 39 @# られ# れ# させ# せ# たがら# E8/E8@ @
```

```
ような@WP 59 @# た# I/EI@ @
ず@WP 69 @ E8/E8@ @
まし@WP 49 @# れ# られ# せ# させ# たがり# E9/E9@ @
だろ@WP 28 @ EIE+MIM+1 @ @
せ@WP 7# @ E8/E8@ @
べき@WP 8I @ EI/EI@ @
たい@WP 9+ @ E9/E9@ @
な@WP4  @ EH/EH@ @
```

付録４：品詞接続テーブル(POSTBL2.TBL,297ﾊﾞｲﾄ)

```
@ Y   @#か# さ# ぞ# ね# よ#  H +@ @
@ 1   @ M 4A 1 @ @
@ E   @#て# ては# ても# M9M#E9EI@ @
@ M   @ @ @
@ C   @ @ @
@ D   @ @ @
@ B   @ @ @
@ A   @ @ @
@ 4   @ @ @
@ R   @ @ @
@ P   @ @ @
@ X   @ @ @
```

付録５：助詞助動詞接続チェック用テーブル(PRTSTR.TBL,6391バイト)

```
          ┌1 台─┐┌─10台─┐┌─20台─┐┌─30台─┐┌─40台─┐┌─50台─┐┌─60台─┐
          1234567890123456789012345678901234567890123456789012345678901234567890
らしい            1                  1      2           111
そうだ                                          1 1     121
う                 1                        2   1 1 1   221           1        2
よう                                11      2        1 121           1         1
る                             1                        2
ようだ                                    2            2
ようです                                  2       1    2
れれ                                                                     2
なけれ                                                                   2
べき                          2211                 1  1     2                2
ような                                    2          1
たろ                2
の                 1 1          222112211 2 2      21   22 2  21     122 1   1 2
から                         1  2111 2      2       1  22     2      122 2   1 1
に                                1         22         11      12     22221  1 1
と                                        2       1  2 2 2   111   1 2111 2  2
を                                        2           1             2    1
が                                        2     1 1
へ                                                   22             22
より                                                                22
で         11                               2       1  2  1           122 2
など                          1 1                    22222112      222 1    1
だけ                          1 1  1      2     1    212 2 12      2 212
まで                          1           2          212 2  2      222 1
か                            1    1      2    11 1  221 2  1      122    1
ばかり                 2        1 1  1            1    2 1 1  2      1 1      2
ほど                          1 1  1                 2 1 1  1      111
ぐらい                             1                 1 11   1 2    1 1
くらい                        1    1             1   1     11 2    1
れ        1122 112 2  2212      1            2     1  2                    2  2
せ        1 211111 2   1 2  2              1 22    1 1                       1
られ      1221 1 1 2  1212                   2                        1       1
させ        1 11   1     1 1                 2
だろ                2
でしょ              2
ましょ              2
ませ           1 12
なく                                         2
んかっ                   2
だっ                     211                 1
でし                     2
まし                     2 2                 1
ように                                    2           1             111    1
ようで                                    2                          1
れる              11          1 1    2   1       1    2221  1111 1 22       1 1
た                11        22  1 2 12221212  2221211 22221  12 2 122       1 2
```

	1台									10台									20台									30台									40台									50台									60台								
	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0
られる																					1				1		1			1				1								1				1	1	2	1					1					1	1										1
せる																															1			1	1								1					2	2	1				1	1						1	1								1		1
させる																																																	1									1													1
ない									1	1		1								1				2		1		1	2	1			1	1	2				2		1	1	1			2	2	2	1				2					1	2			1								2
ず																													1							2											1	1	1					1							2											
だ																											1				1				1	2				2	1	2	1	2			1	2	2	2				1							1									1
ん																									2	1	1	1	2	2							2				1		1	2	1			1	1	2	1				2																		2
です																																					2				2		1	2	2				1	2	2											1												2
ます																																					2							2	1				2	2	2											2												2
ぬ										1																						1				2												2						1		1			1												
ざる																																																	1			2																				
たい																								1							1				2				1		1			1			2	1												1											
も																																					2											2	2																						
は																													1							2									1			1	1																							
でも																																					2											1																								
さえ																																					2																										2									
て			1		1								2		1																					2	1					1	1	1			2	1		2				1	1	2						2	2	1	1	1					1
ので																																					2								1																				2								
ながら																																					2										1																2									
のに																																					2						1																				1								
たり																																					2										1															1	1									
な							1																													2			2									2													2											
よ																																					2											1																								
もの																									1																				1			1				1			2																	
こそ																																																		1													1								
END																																																																						

横軸の助詞・助動詞列

1:なかっ；2:なく；3:ない；4:なけれ；5:られ；6:ず；7:ざる；8:ぬ；9:ん；10: よう；11: う；12: ましょ；13: まし；14: ます；15: ませ；16: た；17: たろ；18: たら；19: たい；20: らしい；21: らしく；22: べき；23: だ；24: だろ；25: だっ；26: でしょ；27: でし；28: です；29: ように；30: ようで；31: ようだ；32: ようです；33: ような；34: そうだ；35: END；36: て；37: たり；38: のに；39: し；40: とか；41: な；42: よ；43: ね；44: もの；45: ても；46: の；47: と；48: が；49: から；50: を；51: へ；52: より；53: で；54: だけ；55: まで；56: ばかり；57: 位；58: ほど；59: ので；60: に；61: も；62: は；63: しか；64: でも；65: さえ；66: ば；67: ども；68: など；69: ながら；70: か；

「END」は助詞・助動詞連続の終了を示す。

（国立国語研究所「電子計算機による新聞の語彙調査Ⅲ」（国立国語研究所報告42，1972，秀英出版）参照）