1969

語彙調査と電子計算機

国立国語研究所

はじめに

国立国語研究所は、昭和23年に設立されて以来、国語や国民の言語生活に関し、諸方面の調査研究を進めてきました。中でも、開所以来つづけている大きな仕事は、現代語の語彙調査です。これまで3回にわたって雑誌の語彙を調べ、それぞれの結果を世に報告しました。これらの調査結果は、現代の国語を知るための資料として各方面で役立っていますが、これからの国語の動向を知るためには、さらに新しい材料によっていっそう大規模な調査を短期間に行なう必要があります。これまでの調査は、すべてカードと人手で行なってきましたので、数十万語の調査に数年を要しました。これを数百万語の規模に上げて、しかも期間をかけないためには、電子計算機を使って作業をするのがもっともよいと考えられます。そこで、昭和40年度に、電子計算機HITAC3010の組織を導入して、200万語の新聞語彙調査を始めました。漢字の入出力という点に大きな困難があり、作業は、はじめ予期したほどの速さでは進みませんが、現状の中でできるだけの方策を考えながら作業を進めています。電子計算機による言語処理という新しい課題の中で、情報処理の基底にある言語処理の問題や言語研究上の新しい問題などに次々に出会います。わたくしたちは、語彙調査作業の中でそれらの問題を研究し、語彙調査プログラムを次第に改善するとともに、それらの問題解決が情報処理一般の進歩にも寄与することを願っています。

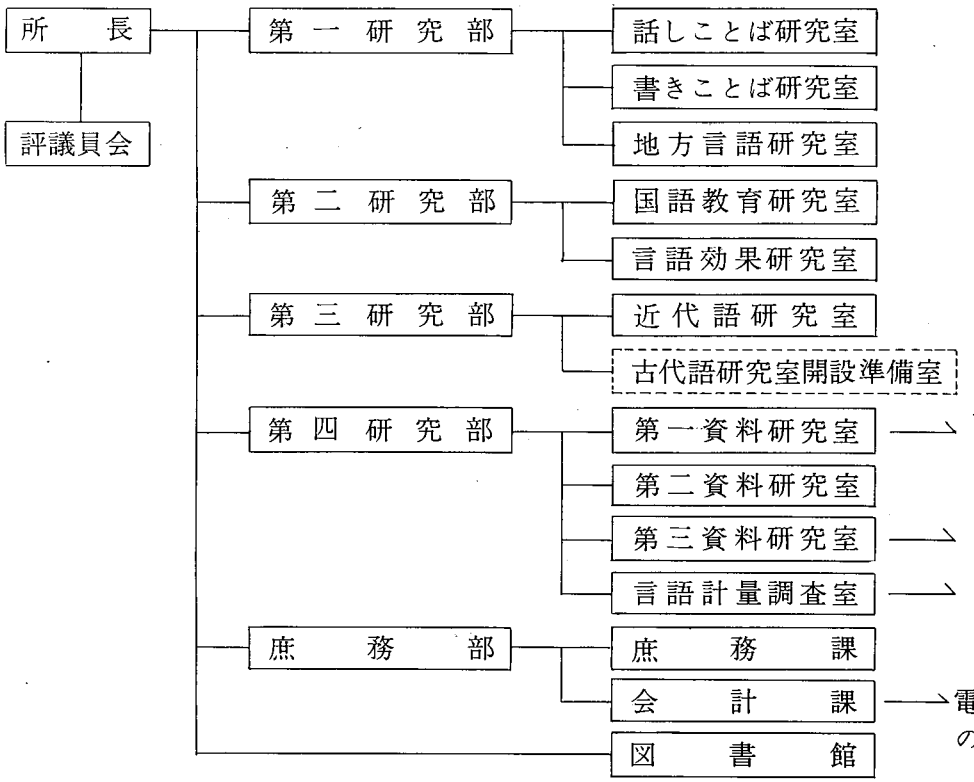
電子計算機組織の導入に際し、諸先達から多くのご指導とお励ましをいただきましたことにつき、あつくお礼を申し上げます。ゆきとどかぬ小冊子ではありますが、わたくしたちの電子計算機による研究業務の概要を報告し、国語研究や情報処理に関心をもつかたがたの導きを得たいと存じます。ご一読のうえ批判・指導のおことばをいただくことができれば幸と存じます。

昭和43年2月

国立国語研究所第四研究部長

林　　四　郎

国立国語研究所の組織と、電子計算機による言語処理業務の担当状況



新聞の用語用字調査

- 語彙表作成→語彙研究
- 語表記表作成→表記法研究

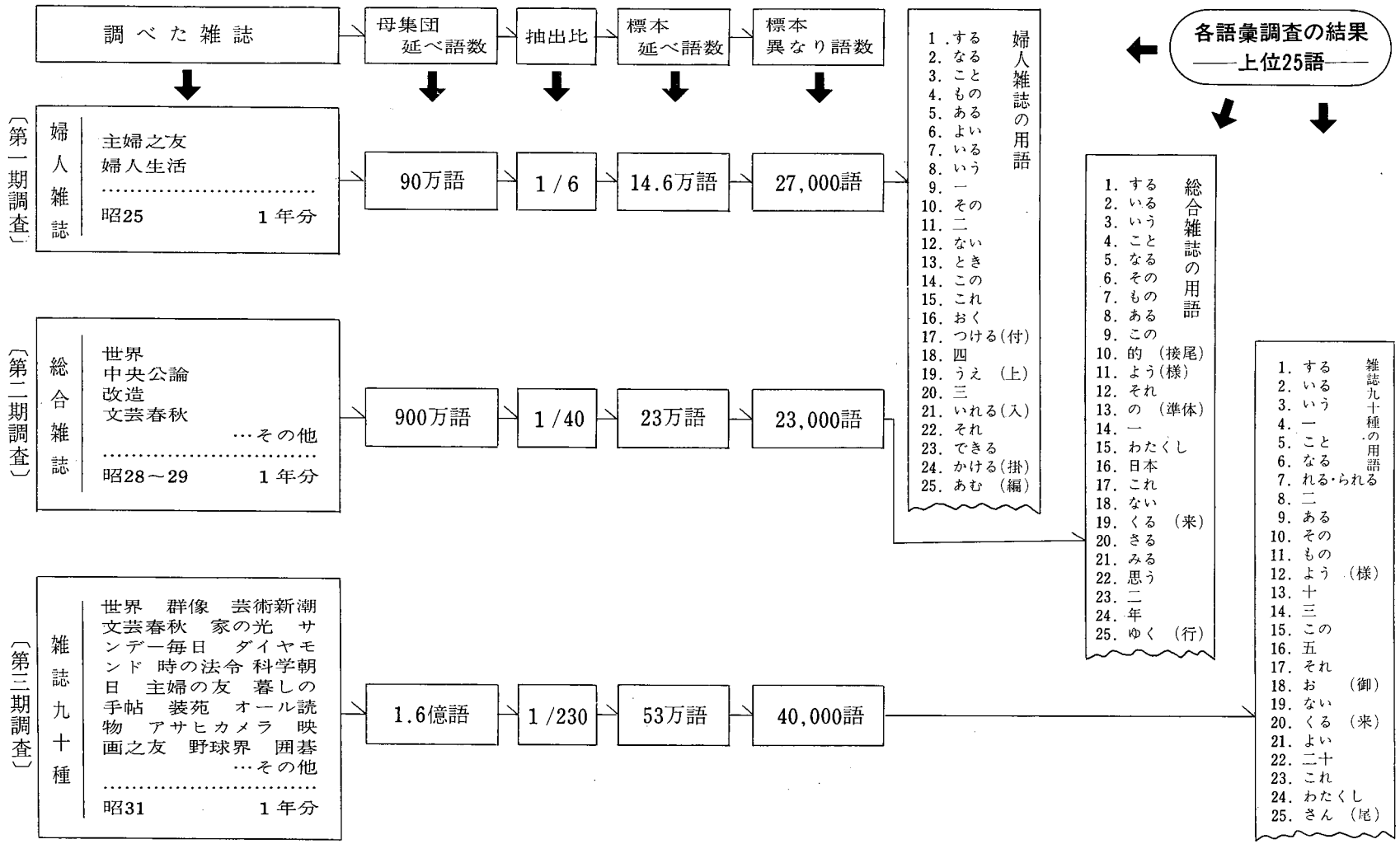
言語データ処理の基礎研究

- 用語総索引作成
- 自動構文分析
- 自動単位分割
- 漢字・かな・ローマ字相互変換システム
- 日本文字のエントロピー
- 録音文字化資料の分析
- 言語行動のシミュレーション

電子計算機　HITAC　3010　昭和41年3月設置　処理装置 H—304(20KC)　磁気テープ装置 H—382　フレキソ H—177
紙テープ読取りセン孔機 H—321　ラインプリンター　H—333C
漢字テレタイプ　沖電気工業製作　昭和40年11月設置

電子計算機を導入する前に国立国語研究所が行なった語彙調査

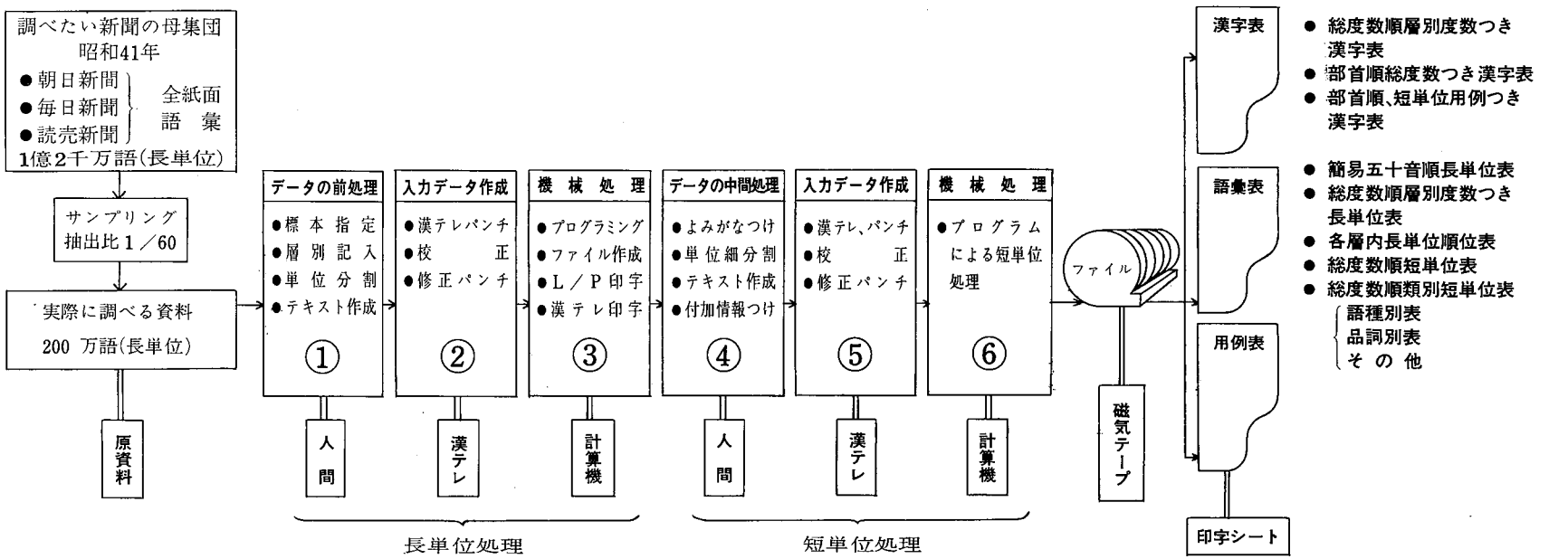
■わたしたちが日常接している読み物には、どのくらいの範囲の単語が、それぞれ、どのくらいの程度で用いられているのだろうか。
■最初に新聞によって、小規模に、調査法を求めるための調査を行ない、以後、雑誌の語彙を調査した。
■どの調査も、ランダム・サンプリングによる標本調査である。第1研究部の書きことば研究室が調査を担当した。



電子計算機による新聞の用語用字調査

■これまでに、雑誌の語彙を調査してきたので、次に、新聞の語彙を調べたい。

■語彙調査のような、量的かつ機械的な業務は、電子計算機を活用して行なうのがもっとも合理的である。昭和40年度から、HITAC3 010を設置し作業を機械処理方式に改めた。第4研究部の言語計量調査室、第1資料研究室、第3資料研究室が担当する。



長単位・短単位とは？

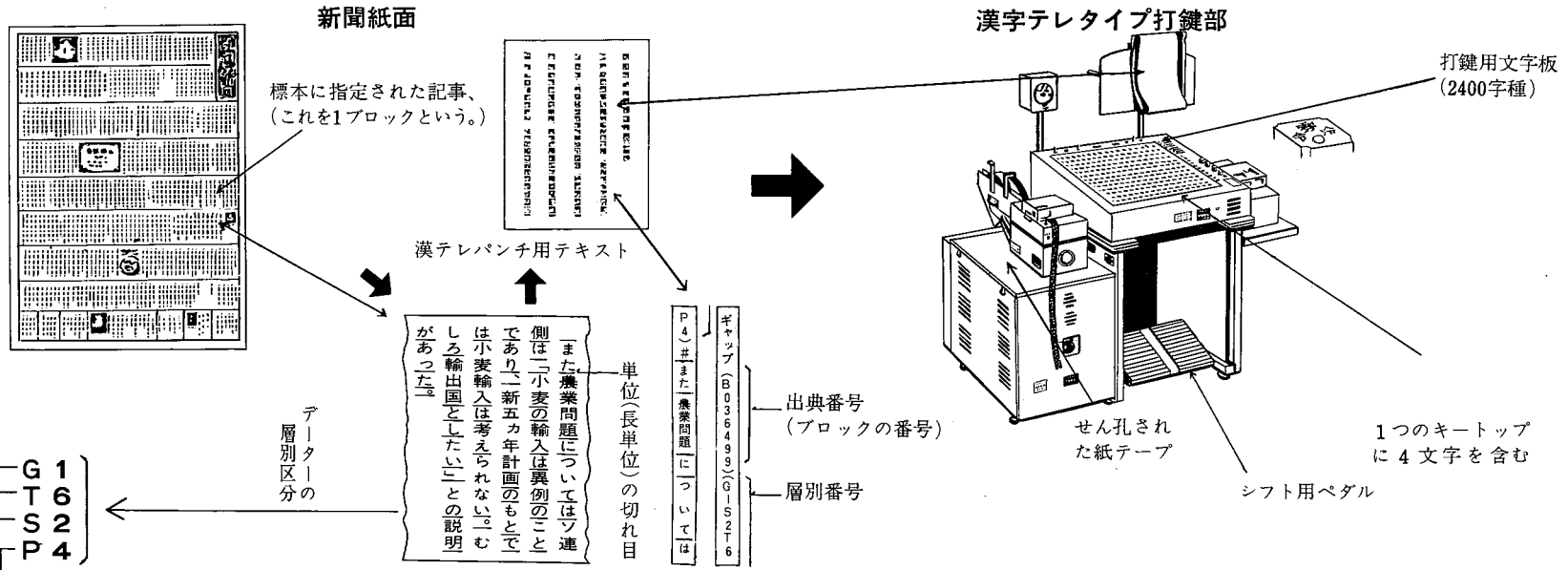
日本語では単語同士が複合して長い熟語を作りやすいので、欧米の諸言語に比べて単語の切れ目を認定するのがむずかしい。そこで、語彙調査をするときには、まず、単語認定の作業基準を作らなければならない。われわれは、長短2種の単位基準を立てた。次の文例で、太線で示したのは長単位、細線で示したのは短単位の切れ目である。

北爆|停止|後|の|ベトナム|問題|は|、……

雑誌の語彙調査のとき、婦人雑誌の調査では長単位で処理し、総合雑誌と九十種とでは、短単位で処理した。現在の新聞の調査では、一度長単位で処理したのち、再度短単位で処理する。

以下のページで、①から⑥までの各処理過程を説明する。

①データの前処理



G(文種)区分

1. ニュース
2. ニュースの解説
3. 社説・コラム
4. ニュース的特集記事
5. 特別読物
6. 評論・論文
7. 実用知識読物
8. 探訪ルポ
9. 長期ニュース展望
10. 記録・通知
11. 紹介記事
12. 読者作文
13. 相互通信
14. 小説
15. 一般広告
16. 案内広告
17. 漫画

T(話題)区分

1. 政治
2. 外交
3. 経済
4. 労働
5. 社会
6. 国際
7. 文化
8. 地方
9. スポーツ
10. 婦人・家庭
11. 芸能・娯楽
12. 広告(G14～G17)

S(署名態度)区分

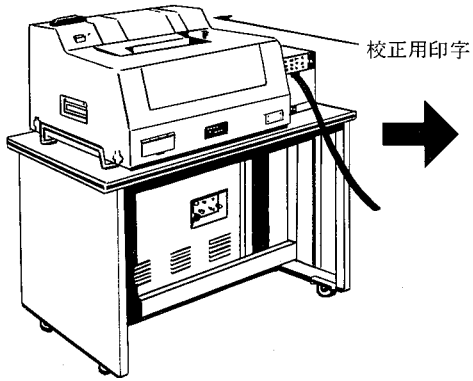
1. 無署名記事
2. 通信社記事
3. 冒頭署名記事
4. 外部者末尾署名記事
5. 記者末尾署名記事
6. 略称末尾署名記事
7. 外電冒頭記名記事
8. 社代表無署名記事
9. 外部無署名記事
10. T12に同じ

P(位置)区分

1. 見出し
2. 標題・欄名
3. リード
4. 本文
5. 情報源・署名
6. 表
7. 写真・図表等の説明
8. T12に同じ

②入力データの作成

漢字テレタイプ印字部

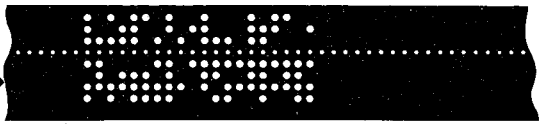


校正作業

毎日〔朝〕3月15日〔火〕

(B036499)(GIT6S2P4) ＃　また
「小麦の輸入は異例のことでは小麦輸入は考えられない
輸　口

修正パンチ



〔ミス・データを含まない正しいデータ〕

2列の孔（1列6ビットゆえ、2列で12ビット）で、漢テレ盤面上の1字を表わす。
この紙テープが、電子計算機への最初の入力データとなる。ただし実際の紙テープはパリティービットを含むので、1列7ビットとなっている。

キートップの文字とペダルによるシフトとの関係

ノンシフト（ペダルを踏まず）
↓
下の文字
右シフト（右ペダルを踏む）
↓
右の文字
左シフト（左ペダルを踏む）
↓
左の文字
両シフト（両ペダルを踏む）
↓
上の文字

鍵盤上にない文字の取り扱いかた

盤外漢字は、盤外字マーク(◆)のあとに盤内漢字2字を組み合わせた3文字によって表わす。組み合わせかたを規定した漢テレコードブックが作ってある。

〔例〕

お釈迦様
⬇
お釈◆定町様

国立国語研究所の漢字テレタイプに収容された文字・符号

当用漢字1845

朕、璽、賑、壱、弐の5字を除いた残りの当用漢字

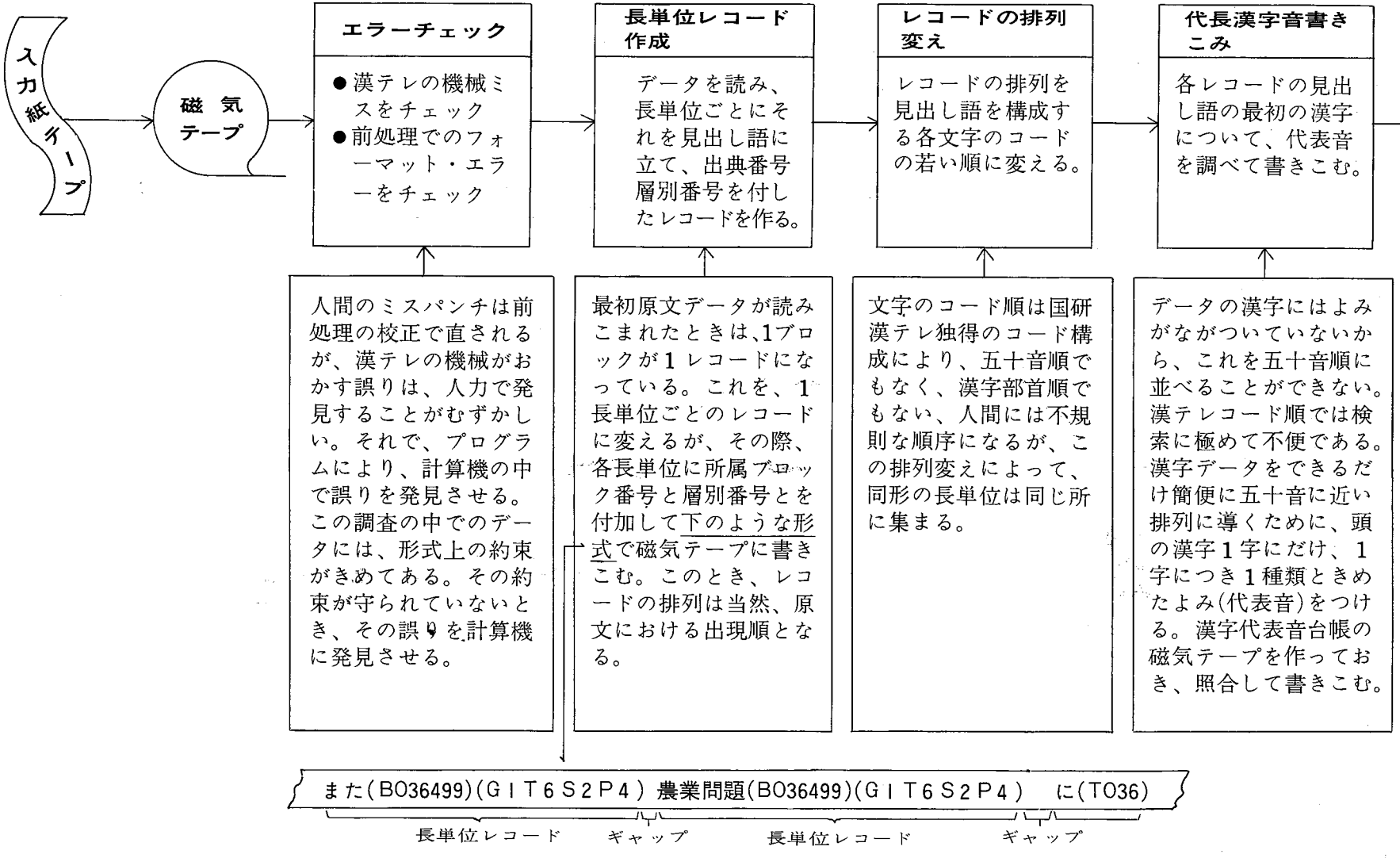
表外漢字 264

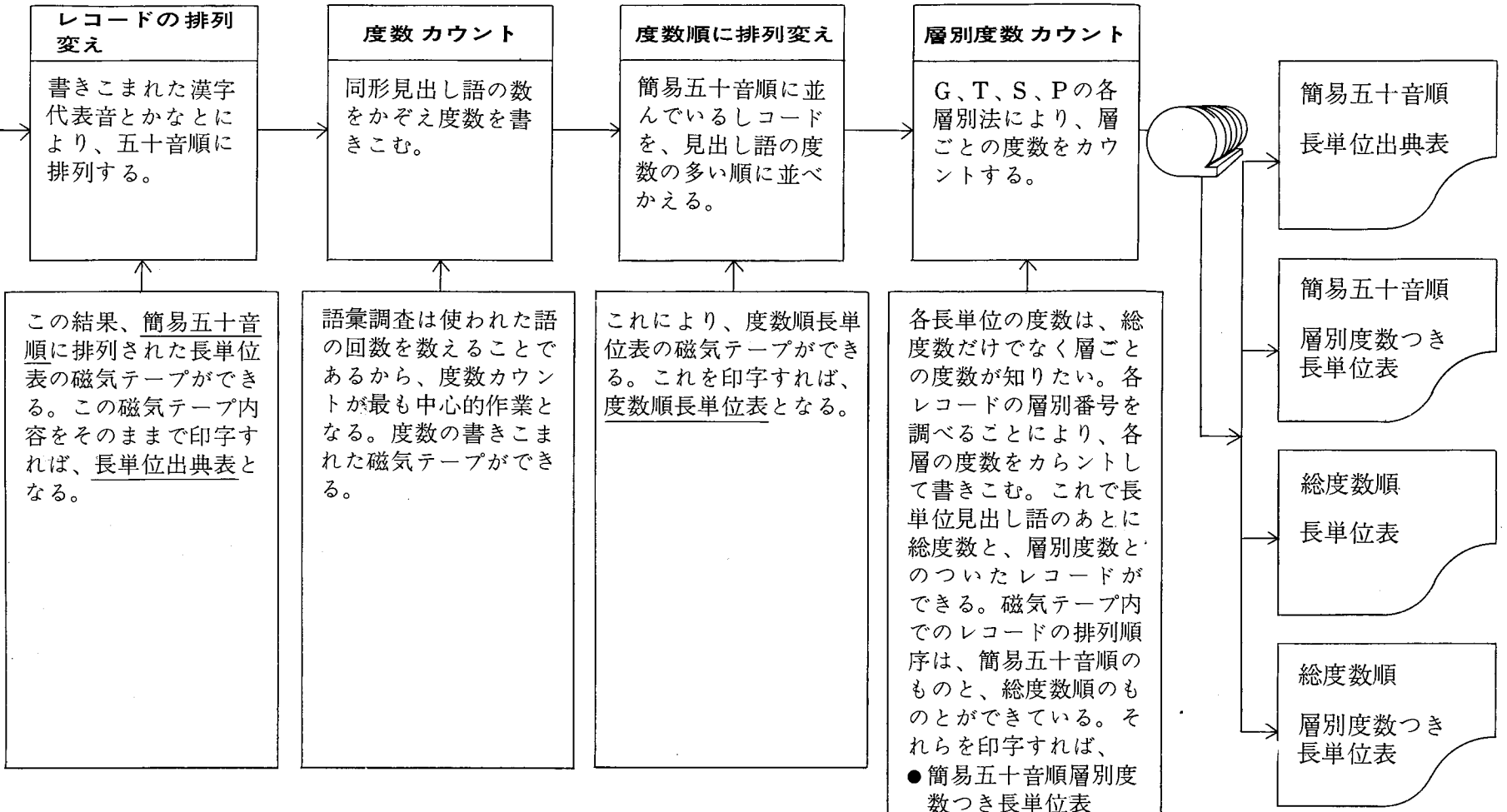
雑誌90種の調査で度数が11以上あったもの

○，々，◆	3
平がな、片かな	170
ローマ字	52
ギリシャ文字 音声記号	16
アラビア数字	10
符号	40
計	2400字

③用語用字調査機械処理プログラムの主要なステップ　I

—最初の入力から長単位表作成まで—

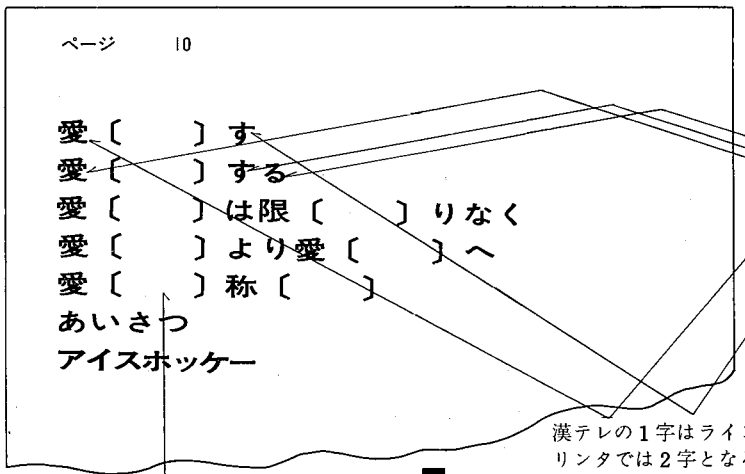



レコードの排列変え
書きこまれた漢字代表音とかなとにより、五十音順に排列する。
この結果、簡易五十音順に排列された長単位表の磁気テープができる。この磁気テープ内容をそのままで印字すれば、長単位出典表となる。
度数カウント
同形見出し語の数をかぞえ度数を書きこむ。
語彙調査は使われた語の回数を数えることであるから、度数カウントが最も中心的作業となる。度数の書きこまれた磁気テープができる。
度数順に排列変え
簡易五十音順に並んでいるしコードを、見出し語の度数の多い順に並べかえる。
これにより、度数順長単位表の磁気テープができる。これを印字すれば、度数順長単位表となる。
層別度数カウント
G、T、S、Pの各層別法により、層ごとの度数をカウントする。
各長単位の度数は、総度数だけでなく層ごとの度数が知りたい。各レコードの層別番号を調べることにより、各層の度数をカラントして書きこむ。これで長単位見出し語のあとに総度数と、層別度数とのついたレコードができる。磁気テープ内でのレコードの排列順序は、簡易五十音順のものと、総度数順のものとができている。それらを印字すれば、
● 簡易五十音順層別度数つき長単位表
● 総度数順層別度数つき長単位表
ができる。
簡易五十音順
長単位出典表
簡易五十音順
層別度数つき
長単位表
総度数順
長単位表
総度数順
層別度数つき
長単位表

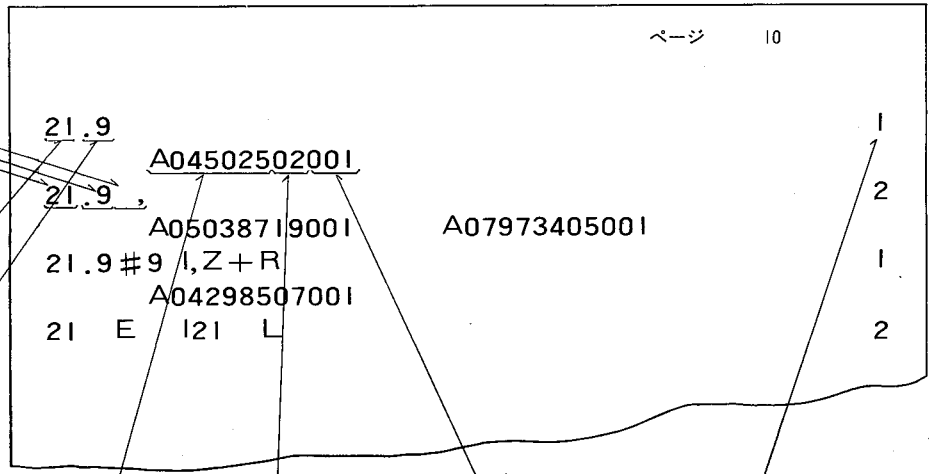
④データの中間処理

簡易五十音順長単位出典表

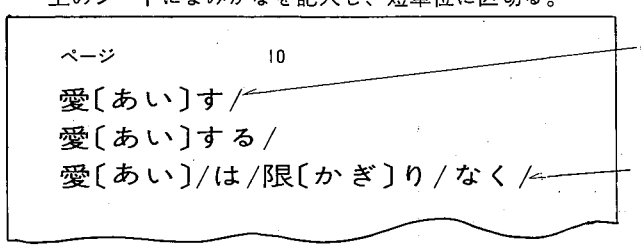
見出し語（漢テレで印字する）



出典表（ライン・プリンタで印字する）



上のシートによみがなを記入し、短単位に区切る。



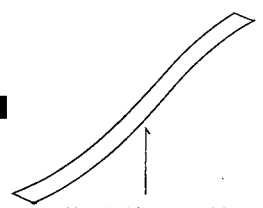
「愛は限りなく」のような連語が長単位で1単位とされるのは、劇などの題名は1長単位とするという、われわれの設けた規定による。
短単位では、このように分割される。

⑥短単位機械処理

ファイル

HITAC 3010

⑤再入力用短単位データ作成



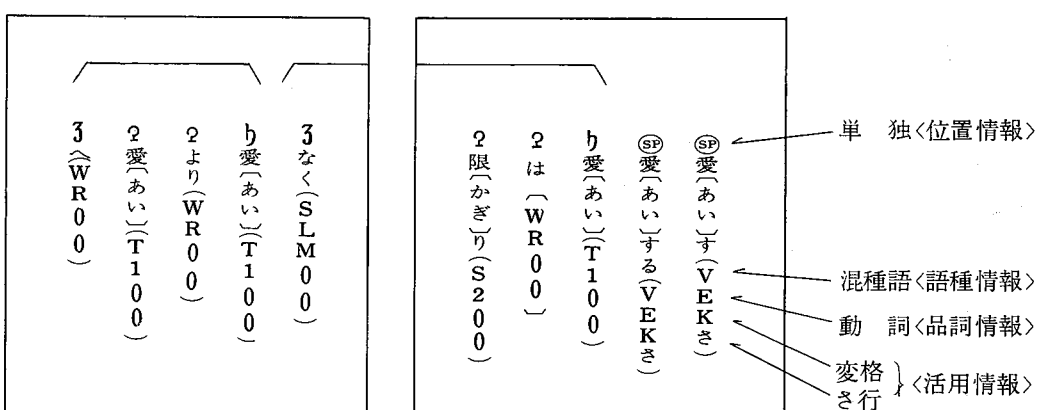
右ページの原稿が打鍵され、紙テープとなったもの。

漢テレ

付加情報コード表

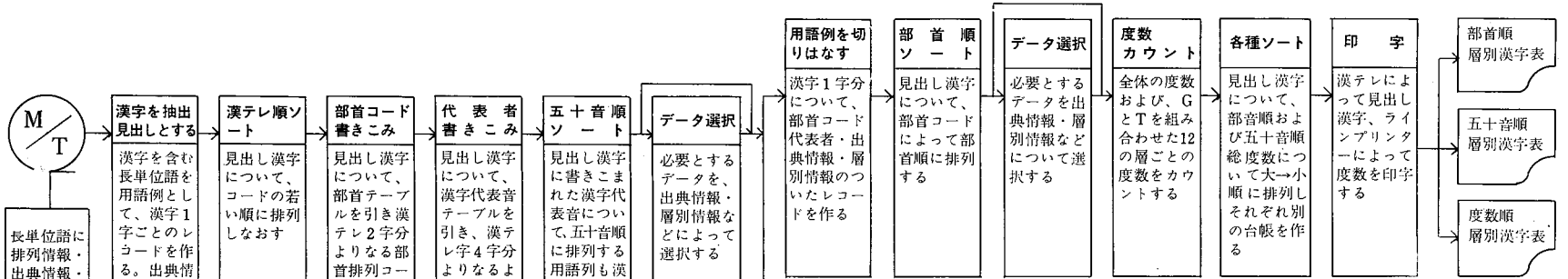
位置情報コード	語種コード	品詞コード		活用コード	
㊎ 単独	S 和語	1 純名詞	D 連体詞	O 無活用	O 動詞以外
ｂ 前部分	T 漢語	2 連用形転成	E 動詞	F 四段・五段	ワ わ・あ行
２ 中部分	U 外来語	3 サ変語幹	＋ 動詞性接辞	G 上一段	あ あ行
３ 後部分	V 混種語	4 形動名	－ 形容詞性接辞	H 上二段	か か行
％ 情報無税	W 語種不要	5 形容名	L 形容詞	I 下一段	さ さ行
	X 数字	6 非用言的接辞	P 助動詞	J 下二段	た た行
	Y 記号	7 数詞	R 助詞	K 変格	な な行
	Z 語種不明	8 固有名詞	X 算用・ローマ数字	M 口語形容詞	は は行
	％ 情報無視	9 代名詞	Y 記号・符号	N 文語形容詞	ま ま行
		A 接続詞	Z 品詞不明	P 助動詞	や や行
		B 感動詞	％ 情報無視	Q 形動語尾	ら ら行
		C 副詞		％ 情報無視	わ わ行
					％ 情報無視

よみがなの記入され、短単位に切られた見出し語のシートと 上の
付加情報コード表 とから下の再入力用挨テレ・テキストを作る

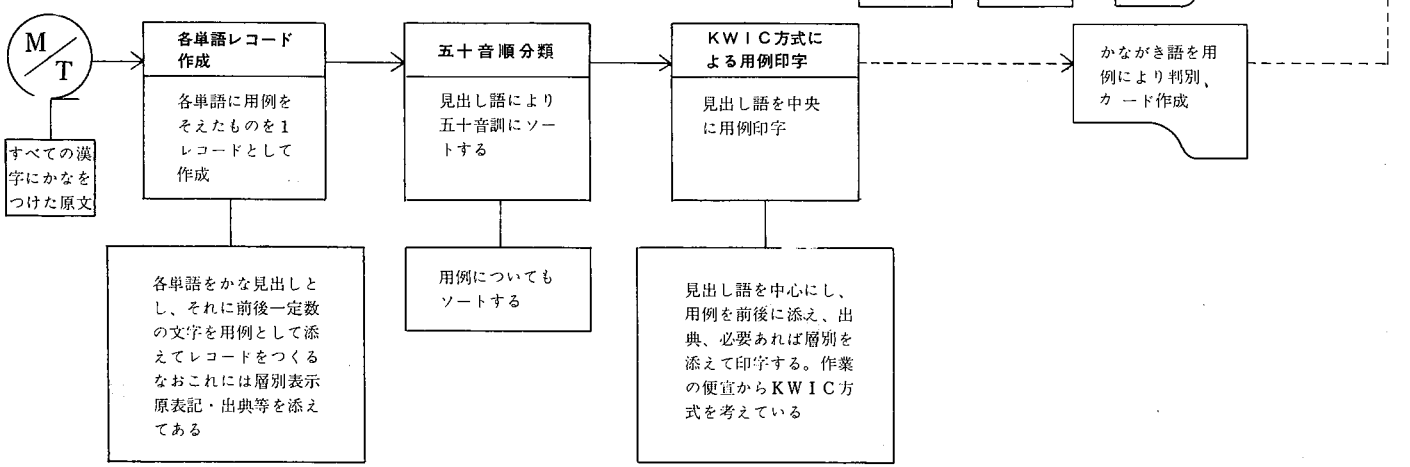


③用語・用字調査機械処理プログラムの主要なステップII

〈開発ずみ〉—長単位表ファイルから漢字表作成まで—

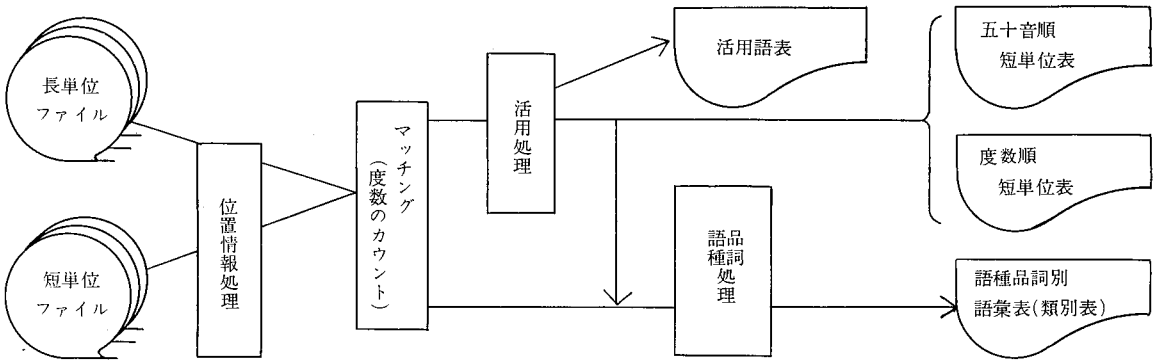


《設計中》—短単位表ファイルから漢字表作成まで—



⑥用語・用字調査機械処理プログラムの主要なステップ　III

—短単位データ入力から短単位表作成まで—



度数	配列情報	短単位 よみがな	短単位 漢テレ	INDEX	付加情報 個数	語種	品詞	活用		位置	長単位 よみがな	長単位 漢テレ	よみがな付短単位
		かぎり	限り		1	S	2	0	0	2	あいはかぎりなく	愛は限りなく	限〔かぎ〕り．

度数	全体順位	全体比率	部分順位	部分比率	見出し語 配列情報	見出し語 よみがな	見出し語 漢テレ	INDEX	付加情報 個数	語種	品詞	活用		見出し語 よみがなつき
						かぎり	限り		1	S	2	0	0	限〔かぎ〕り

度数	配列情報	代表形 よみがな	代表形 漢テレ	短単位 よみがな	短単位 漢テレ	INDEX	付加情報 活用形	個数	語種	品詞	活用		見出し語 よみがな付短単位
		あがる	上がる	あがり	上がり		連用形	1	S	E	F	ら	上〔あ〕がり

1．活用語彙表

各活用語について、代表形（終止形）と度数を示し変化形別の度数カウントを入れる。

2．五十音順短単位

各見出し語別に、語種・品詞・活用コードと度数順位を示す。

3．度数順短単位表

各見出し語別に、語種・品詞・活用コードと度数使用率等を示す

4．語種品詞別語彙表(類別表)

各見出し語について度数、類内順位、類内使用率等を示す

語彙調査以外の言語処理

研究と成果

用語総索引作成〈プログラム作成ずみ〉

A. 漢字かなまじり表記によるもの　漢テレを入出力とし、分かち書きのしてある漢字かなまじり文データをそのまま扱って、見出し語のもとに所属センテンスをすべて印字する。

B. かなまたはローマ字表記によるもの　分かち書きのしてあるかなまたはローマ字表記のデータを扱って、次のようにアウトプットする。

1) センテンスごとに改行した原文
2) 見出し語（五十音順またはアルファベット順）に出典情報（所属段落、段落内の文番号、文内の語番号）をそえたもの
3) KWIC方式による用語一覧表

自動単位切り〈研究中〉

分かち書きのしてない日本文を計算機に読ませ、計算機自身に単位切りを行なわせることを目標とする。日本文データを扱うときには、この問題はつねに足元にあるので、この課題は、国語データ処理における出発点である。そして、この課題の解決は、日本文の自動解析にまで進むので、国語データ処理の帰結点でもある。われわれは現在、次の二つの方式で問題に迫っている。

1) 漢字とかなの排列上の連続確率から、切れつづきを機械的に推定する方式。これは方法が簡便で処理時間が短かい。
2) 1センテンス内の文字連続を、辞書の照合によって語と認定し、過不足なく認定できたら、その認定を正しいものとする。この認定のために文法規則が使われるので、結局文の自動解析をすることになる。可能な解析結果はすべてアウトプットする。この方法は辞書と文法規則を必要とし、かつ照合回数が多大となるので、処理に時間がかかり回り遠い方法だが、根本的な解決法である。

漢字－かな
漢字－ローマ字　相互変換システム〈実験中〉

漢テレを使用すると、入出力に大量の労力と時間を要するので、計算機の処理スピードと大きなへだたりを生じ、能率がきわめてわるい。計算機自体が漢字をかなやローマ字に直したり、反対に、かなやローマ字を漢字に直したりできれば日本語のデータを処理するのにたいへん便利である。このような相互変換システムを目標にして、まず、漢字をかなに変換するシステムを作成している。国研漢テレの盤内にある漢字2110字の音訓を辞書として磁気テープに入れておき、計算機に音訓の選び方を文脈から推定する方法を教え選ばせるプログラムを作って実験中である。現在のプログラムで新聞の文章を処理した場合、成功率は固有名詞をのぞき80%程度である。

漢字かなまじり文における文字のエントロピー〈プログラム作成ずみ〉

漢字かなまじり文一般において、ある文字の次にどういう文字が来るか、ある文字連続のあとにどういう文字が来るかを予測させる文字連続の確率を計算する。この計算は、極めて大量の文字データについて計算しなければ意味がない。これまで、そのようなデータがなかったが、現在われわれが行なっている新聞語彙調査は、データ量から見て今までにない分量のものである。この量もエントロピーの計算のために決して充分なものではないが、とりあえず材料とするに足りる。

言語行動の語彙論的・場面論的分析〈プログラム作成ずみ〉

分かち書きのされたローマ字日本文データに形態音韻論的分析を施したものを入力データとし、これを、発話者、発話目的や場面、コミュニケーションの相手等によって、用語を分類集計するプログラム。最初の資料として昭和38年に松江市で行なった「国民各層の言語生活の実態調査」のうちの、一家族の24時間調査録音テープを用い、分析した結果が近く発表される。

言語行動のシミュレーション〈研究中〉

人間の言語行動をシミュレートすることはむずかしい問題であるが、言語を行動として見ようとする立場からは、これは重要な研究テーマであり、有効な方法である。いろいろな方法があろうが、文の構造や文章の構造を自動的に解析するアルゴリズムが作られれば、言語行動のシミュレーションの一つと見ることができよう。

電子計算機による言語データ処理の展望

電子計算機による言語の各種の処理を一般に言語情報処理(Language Data Processing)といっている。このなかには、言語研究における基礎的研究の段階にあるものもあるが、すでにかなり一般的に実用化したものもあり、他の情報処理の一部分として使用されているものもある。言語データ処理は、全体として未開拓な分野が多く、特に言語自身に関係した部分の基礎的な研究開発が期待されており、言語学の新分野として計算言語学(Computational Linguistics)が誕生した。前ページまでに記した国立国語研究所の諸研究はいずれも計算言語学の中にかぞえられるものだが、このページには、国語研究所では現在行なっていないが、他の各所で行なわれているもののうち、重要なテーマのいくつかをかかげた。

言語研究　Lingnistics

用語総索引の作成、用語調査、構文解析、文章解析など、言語研究それ自体のために電子計算機が広く用いられている。

機械翻訳　Machine Translation

ある言語から他の言語へと、機械を用いて自動的に翻訳すること。自動翻訳ともいう。言語情報処理の中では、アメリカ、イギリス、イタリー、日本などで早くから研究され、一般によく知られている分野であるが、言語自動分析の困難から早急に実用化される見通しはうすい。しかし、人間が前処理(pre-edit)、後処理(post-edit)をしたり、人間翻訳の手伝いをさせる半機械翻訳(machine aided translation)の面で実用化を期待する向きもある。

自動抄録、自動索引づけ　Automatic Abstracting Automatic Indexing

論文の内容を計算機に抄録させたり、記録内容を端的に示すkey wordsを抽出選定させたりすること。従来、人間がこの種の作業を行なってきたが、作業の均質化、将来のコストの点などから、人手にばかり頼れず、機械による自動作業化の必要が痛感されてきた。方法としては、まず用語調査を行なって、用語の使用頻度を手がかりとする方法と、自動構文解析(automatic syntactic analysis)や自動文章解析(automatic discourse analysis)によって論理的に解析する方法とがある。一般には、前者が簡便なのでよく用いられるが、本質的な解決は後者にまたなければならない。

新聞・速報等の自動編集

文書・報道などを自動的に編集し、組版するシステムが開発され、実用化されている。アメリカでは新聞編集で、日本では科学技術情報センターの研究文献要約速報などで現に実用化している。

情報検索　Information Retrieval

現代は情報の洪水時代といわれ、大量の情報を集めておき、必要な場合に必要な情報が抜き出せることが望まれる。この作業の機械化が情報検索である。検索にあたっては、必要な情報をもれなく抜き出すために、当該分野での同義語集、類義語集、関連語集、上位・下位概念語集（シソーラス・thesaurus）が使われることが多い。また、抄録や索引づけは当然付随するもので、これが一体となった情報検索の自動化の研究が進められている。例えば、アメリカにSMART方式(Salton's Magical Automatic Retrieval of Taxasの略)と呼ばれるシステムがある。

パタン認識　Pattern Recognition

現在、計算機への入力には、いったんキーパンチャーの手によって紙テープかカードにさん孔されたものを用いるのがふつうで、光学読取(OCR)も一部で用いられているが、文字の種類等その使用範囲は極めて限られている。活字の漢字あるいは手書きの文字の読み取りができるようになれば、その利益ははかり知れない。また、音声による機械との会話（音声の認識と合成、音声タイプライタ）ができれば、一般事務の能率化はさらに急速に進むであろう。まだ研究の初期にあるが、実験室の段階では、すでに実現を見つつある。

人間と機械との通話　Man-machine Communication

プログラム言語を人間の言語に近づけて、人間が人間に対するように、ふつうのことばで質問や命令を行なうと、計算機がそれに答えたり、命令を実行したりするようになることを目標とした研究。また、計算機の記憶容量の大量化、処理のスピード・アップによって多数の端末機器からの入力および、それへの出力が可能になってきたので、質問に対する答えが時分割で同時化されつつある。その結果、大組織の中央一括管理や、教育における多数指導と個別指導との両立に道が開けて来た。

その他

教育のプログラム化、一般事務の能率化などにも、言語データ処理の問題が内在する。例えば電話帳の編集、住民台帳や名簿の作成など、いろいろな業務が電子計算機によるデータ処理の開発を待っている。