

コンピュータによる日本語情報処理

The EDP System for Japanese Information Processing

Like image information processing and data base systems the Japanese character information system is regarded as one of the most important tools in ushering in a new era of computer utility in Japan. Its input system, however, is tackling presently a very difficult problem of handling a vast number of Japanese characters. Although the full key board input system is in general use because of its outside character-set processing capability, such new systems as "Kana" typewriter input system and handwritten Japanese character recognition system which are still under development are much hoped for from the reason of operational facility. One of the important themes for processing systems is the language translation from phonetic character to ideograph and the development of a question and answering system by natural language. As for the output system the CTS (Computerized Typesetting System), and address printing system are expected to make rapid development.

This article introduces Japanese information systems centering around Hitachi's achievements in this field.

藤本 謹也* *Kinya Fujimoto*

鎌田 素明* *Motoaki Kamata*

小田切郁夫* *Ikuo Odagiri*

1 緒 言

コンピュータ システムの利用形態が技術計算や大量統計処理から進んで、さらに高度な情報処理、データ ベース システムへと進むにつれ、マン・マシン間のコミュニケーション手段として日本語情報処理とイメージ情報処理の必要が生じてきた。すなわち、コンピュータの取り扱う情報が数字やかな文字から漢字かな交じり文の日本語やイメージ情報へと発展が要求されている。日本語情報処理の適用分野は、現在日本語文書を使用しているほとんどの分野であり、行政管理庁の調査によると、ほとんどの省庁が将来日本語情報処理システムの導入を検討している。日本語情報処理は他の情報処理技術に伴う情報の収集、蓄積、加工、検索などのシステムとの総合的関連で取り上げられるため、今後日本語を取り扱わないコンピュータは存続不可能といわれるほど重要視されてきている。

2 日本語情報処理システム開発の問題点

日本語情報処理システムを論ずるときに重要な問題である日本語、特に漢字自体の特性からくる問題点について述べる。

(1) 漢字は字種が非常に多く、かつ分野により使用字種が異なる。

漢字字種は5万字とも6万字ともいわれているが、現在わが国で一般に用いられている漢字字種は3,000字前後が多い⁽⁴⁾。しかし英語などの他の言語と違って使用字種が多いため、特に入力システムの大きな壁となっており、採用字種の選択がシステム設計上の重要ポイントである。

(2) 字画が複雑である。

(3) 漢字の分類、配列の規準がいくつもある。

漢字の配列には、画数順、部首/画数順(漢和辞典)、音読、または訓読の五十音順・イロハ順などの方式があり、ソートは漢字コードの決め方により決定される。

(4) 漢字は読みが一意に決まらない。

表意文字である漢字は1字で多数の読み方があり、表音文字のかな文字は多数の同音異字を持っている。

(5) 漢字コードが標準化されていない。

そのため、システム間の日本語情報のデータ交換は、字種やコードの相違により現在は不可能に近い状態である。

しかしながら、このような問題を含みながらも、日本語情報処理システムの実現は、しだいにその数を増加し、それぞれに各種の工夫を凝らしている。これらの各種工夫が集積されて、処理の標準化がなされることが必要である。

3 日本語情報処理システムの動向

日本語情報処理システムには多くの利用形態があるが、おもなものは次のとおりである。

(1) 新聞、出版物の編集組版

新聞、印刷業界では鉛や騒音の問題、熟練作業者の不足などで、早い時期からCTS (Computerized Typesetting System) の開発が行なわれている。コンピュータによる編集組版システムは、単に編集の合理化だけでなく、編集情報の蓄積、加工、検索などの多方面の利用が行なわれている。

(2) 宛(あて)名印刷

DMS (Direct Mail Service)、証券代行業、行政官庁などでは漢字による氏名、住所のマスタファイルを作成しておき、必要に応じて宛名の印刷を行なっている。

(3) 日本語情報の表示

放送会社における選挙速報、証券取引所における株価の表示など、時々刻々に変化する情報を受取者の意志に関係なく漢字ディスプレイ装置に表示するシステムである。

(4) 情報検索

日本語データ ベースに漢字、かな文字、または数字コード

* 日立製作所ソフトウェア工場

表1 日本語情報処理システムの利用形態とアプリケーション例
各利用形態とその適用システムの相互関連の強いものを*で示した。

Table 1 The Usage and Applications of Japanese Information Processing System

アプリケーション 利用形態	アプリケーション											
	1	2	3	4	5	6	7	8	9	10	11	12
	ダイ メレクト サービス	印 刷 出 版	速 記 録 作 成 シ ス テ ム	会 議 録 情 報 シ ス テ ム	放 送	新 聞 発 行	登 録 管 理 シ ス テ ム	顧 客 管 理 シ ス テ ム	証 券 代 行	情 報 提 供 サ ー ビ ス	人 事 ・ 統 計	国 語 研 究
1	宛 名 印 刷	*					*	*	*			
2	編 集 ・ 組 版		*	*	*	*				*		
3	日 本 語 表 示				*				*	*		
4	情 報 検 索		*	*	*	*	*	*	*	*	*	*
5	言 語 変 換		*	*						*		*
6	事 務 処 理	*					*	*	*		*	

で検索し各種情報を得る。問合せのレベルには、

- (a) 個人の特定を目的とした漢字氏名による検索
- (b) キーワードを漢字で表現した論理検索
- (c) 自然語（漢字かな交じりの日常語）による情報検索

のように分けられる。

(5) 言語変換

かな文字、ローマ字、速記などの表音文字を入力し、読みやすい漢字かな交じり文に変換したり、既存のかたかなによる氏名 住所マスタを漢字マスタに変換したり、同時通訳、音声入力・音声応答のシステムである。

(6) 一般事務処理

漢字により情報の正確化を図るため人事管理などでは情報の漢字化が行なわれており、一部では契約書、定型文書の自動作成システムが実現している。

表1は、日本語情報処理の利用形態とアプリケーションを示すものである。

4 日本語情報処理システム

日本語情報処理システムの基本的機能は次のように構成される。

- (1) 入力システム
- (2) 処理システム
- (3) 出力システム

以下各機能について、各種システムにおける実績を中心に説明する。

4.1 入力システム

日本語情報処理においては2.で述べた日本語の持つ特殊性から入力システムを一段と複雑化、かつ特殊化している。そのため日本語情報処理システムの運用に要するコスト、所要工数、人数などに対する入力システムが占める割合は非常に大きい。

以下に、入力方式の現状および問題点と国立国会図書館の指導のもとに共同開発した外字処理(文字分解方式)、校正処理、機械辞書について説明する。

4.1.1 入力方式

日本語情報の入力方式には、表2に示すようにすでに商品化されているものおよび研究中のものなど多くの方式があるが、操作性、経済性の面で問題があり、今後、商品化される方式に大きな期待が寄せられている。

入力方式における問題点としては一般的に次のようなものがあげられる。

- (1) 入力コストが高くなること。
- (2) 専門のオペレータが必要で、素人(しろうと)が扱いにくいこと。
- (3) 入力のスピードが遅いこと。
- (4) 入力の精度を向上させるための効率よい手段が確立していないこと。
- (5) 分野により使用文字種の片寄りが大きいため、入力機器の汎用性に乏しいこと。

表2 日本語情報の入力方式一覧
日本語情報の入力には、各方式ともそれぞれ特徴があり、各アプリケーションに最適な入力方式を選択する必要がある。

Table 2 The List of Input Process Pattern of Japanese Information

項番	方 式	概 要	特 徴	備 考
1	フルキーボード方式	○鍵盤のキートップに文字を配置し、シフトキーにより一つのキートップのうちの一つの文字を指定する。 ○4～15段シフトがあり、相当の熟練を要する。	○盲パンチが多く、検孔機能がないものが多い。 ○取扱文字数 1,100～5,400字 ○入力速度 平均60字/分	○漢字テレタイプとして、古くから多く利用されている。
2	数字コード方式	○1文字を4けたの数字としてコード表現して入力する。 ○数字コードの入力は、すべてオペレータの記憶にたよる。 ○フルキーボード方式の外字処理として用いられる。	○装置が安価で検孔も可能である。 ○コンピュータによる内部コード変換が必要である。	○中華人民共和国の電信に用いられている。
3	文字分解入力方式	○文字を字形と字素に分解して入力し、コンピュータにより漢字辞書(テーブル)を参照して漢字コードに変換する。 ○フルキーボード方式の外字処理として用いられる。	○タッチ数が増加する。 ○文字組立てに漢字辞書とソフトウェアが必要である。	○外字処理の項を参照。
4	文字盤指定方式	○平面形または円筒形の文字盤テーブルの所定文字を指定することにより入力できる。 ○和文タイプの改造形のもの磁気ペンなどによる文字指定とがある。	○和文タイプ改造式はモニタがとれる。 ○操作が容易であるが、外字処理が困難である。	○平面形の例：日立製作所小田原工場漢字ディスプレイ端末装置の研究(磁気ペン入力方式)
5	かな鍵盤入力方式(配列対応形)	○平面に漢字が配列されており、所定の漢字の存在するアドレスを「親」文字盤、「子」文字盤の2タッチで指定する。	○装置が簡単で、腕の動きが少ない。 ○漢字配列の記憶、訓練が鍵盤方式より困難である。	○オンライン端末として用いられている。
6	かな鍵盤入力方式(連想記憶形)	○漢字1字を意味や形、熟語などの連想を手だてとして記憶して入力する。 例：ハム→公、Uミ→海、ロロ→回、ボテ→宿、チラ→瞬	○記憶が容易で入力速度が速い。 ○操作が容易であるが、外字処理が困難である。	○ラインプット方式とも呼ぶ。
7	かな鍵盤入力方式(表示選択形)	○漢字の読みをかな文字で入力、その音に対応する漢字をすべてディスプレイ装置に表示して該当する漢字を指定する。	○訓練は不要であるが、入力速度が遅い。 ○ディスプレイ装置、辞書が必要で高価となる。	
8	かな鍵盤入力方式(フィードバック方式)	○漢字の訓読みをかな2文字で入力し、同音異字があるときは機械のほうから追加情報を要求するチャイムを鳴らし、3文字め、つくりなどの情報を入力する。	○盲打ちが可能であり、操作が容易である。 ○慣れると機械からのフィードバックにたよらなくてもよい。	○日立製作所中央研究所かなタイプ入力による邦字、漢字入力装置。
9	かな鍵盤入力方式(ソフト変換方式)	○かな文字で入力してコンピュータの内部で漢字かな交じり文に変換する。	○分かち書き、漢字の指定などの操作が必要である。 ○同音異字の対策が必要である。	
10	速記入力方式	○速記用タイプライタを用いて話しことばを入力し、コンピュータにより漢字かな交じり文に変換する。	○専門オペレータの養成に長期間かかる。 ○同音異字の対策が必要である。	○日立製作所ソフトウェア工場速記自動反訳システム
11	手書き漢字入力方式	○ミニコンピュータとタブレット板を接続し、タブレット上で文字を手書きし、その字画、書き順、パターンの情報を用いて漢字を認識する。	○楷(かい)書体で、書き順を正しく書く必要がある。	○日立製作所中央研究所漢字・かたかな、オンライン手書き文字認識装置
12	印刷漢字認識方式	○印刷されている漢字のパターンを階層的パターンマッチングによる方法で何段階かにぼかしてカテゴライズする。	○印刷漢字の入力に適している。	○日立製作所中央研究所印刷漢字認識入力装置

これらの各種方式のうち、日立製作所では入力迅速性、正確性、操作性、汎用性の向上を目指し新しい各種入力方式を研究開発中である。

4.1.2 外字処理

現在入力機器では装置の価格面、操作面から2,400字程度扱える装置が多い。システムでそれ以上の文字数を必要としていて入力装置にない文字種を外字といい、スムーズな取り扱いが可能となるような配慮がなされている。

機械化対象業務が決定されると使用文字の頻(ひん)度調査を行ない出現頻度順の文字種の一覧表を作成する。最も一般的なフルキーボード方式では頻度表の上位から2,000字、2,400字など入力装置の文字収容能力に合わせた字数を収容する。したがって外字は比較的出現頻度は低いが、システムには欠かせない文字種が該当する。

外字処理の方式として、数字コード方式、文字分解方式(パターン合成法)について説明する。

(1) 数字コード方式

4けたの数字を入力することにより、ソフトウェアで漢字コードに変換する方法である。

例：峠…1274 1274…峠の漢字コード
(コードブック) 入力・ソフトウェア変換 X'57B2'

(2) 文字分解方式

文字を構成する字素に分解し、字素を組み立てている字型を定義することにより、それらのおののを入力し漢字辞書という一種のテーブルを参照して、該当する文字コードに置き換える方法である。このとき鍵(けん)盤に存在する文字を外字処理しても、また別の字型で入力しても同一コードになる必要がある。

例：峠……> 田, 山, 上, 下 峠の漢字コード
(分解) (字型) (字素) (入力) X'57B2'

4.1.3 校正処理

現在広く用いられている入力方式では、入力情報のエラー率は入力方式やオペレータの習熟度などによっても異なるが、だいたい1~10%程度である。経済性、処理速度などから入力精度の向上のための画期的方式のない現在では、入力情報をチェックし、誤りを発見して修正する校正処理が重要である。

校正処理は人手に依存する度合いが高く、校正処理のターンアラウンドタイムが長くなり、日本語情報処理のネックとなっている。

(1) モニタ校正

入力した情報をそのまま漢字プリンタに出力し、入力原稿と読み合わせ、エラー修正情報を再入力し、更新する方式である。

(a) モニタ校正の特徴

- (i) コンピュータとオフラインで修正され、バッチ処理する。
- (ii) 校正情報にエラーが発生するため、校正結果のモニタを取り、再び校正するというサイクルをくり返す。

(b) モニタ校正の例

モニタ校正の例として日立製作所で開発、実施している「日本語文ドキュメント自動組版システム(HDES)」の校正方式について述べる。

- (i) ページ付けしたリストを出力し、ページ番号、行番号を用いて行を指定し、さらにその行の校正すべき個所の前

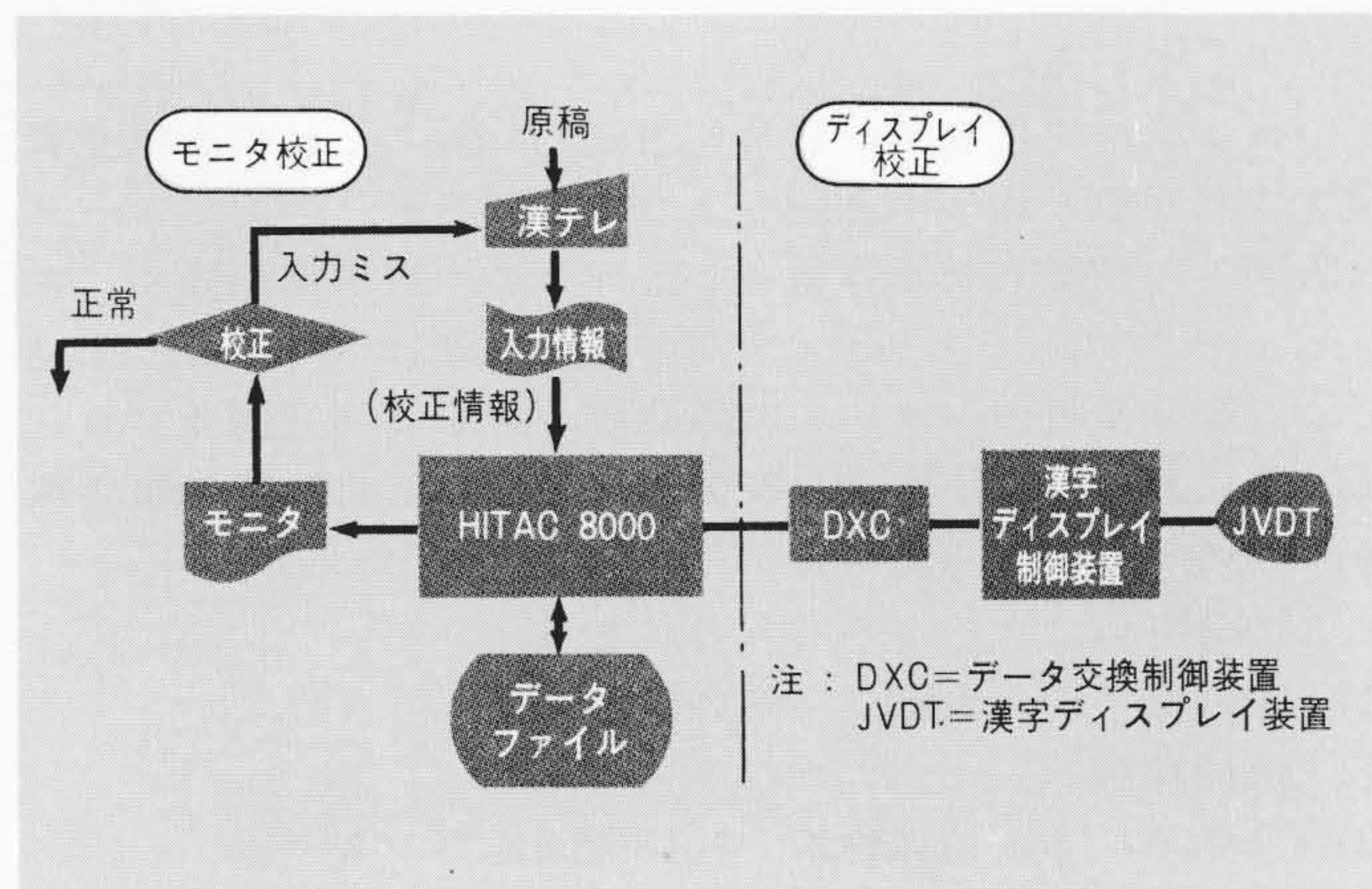


図1 校正処理概念図 モニタ校正とディスプレイ校正を組み合わせ、校正の完全化とターンアラウンドタイムの短縮を図る。

Fig. 1 Error Correction System

後1文字をキーキャラクタとする。

- (ii) 校正情報の精度向上のため、キーキャラクタの指定誤りに対するプロテクト機能を設ける。
- (iii) 校正情報の校正を完全に行なってからマスタの校正を行なう。

(2) ディスプレイ校正

中央処理装置に連結された漢字ディスプレイ装置からランダムファイルに記録された情報を任意に取り出して、ディスプレイ上に表示し修正を行なう方式である。

ディスプレイ校正の特徴は次のとおりである。

- (a) 校正結果が直ちにディスプレイ上で目視確認できるため、校正段階でのミスの再発は防止しやすい。
- (b) 漢字ディスプレイ装置が高価なため、コスト高となる。

そこで第一次校正、第二次校正など比較的错误が多い段階ではモニタ校正を行ない、校正処理が急がれる場合や、エラーが少なくなった段階ではディスプレイ校正を行なって校正の完全化を図るのが得策である。図1は校正処理の概念図を示すものである。

本方式は、国立国会図書館において昭和48年8月より開始したシステムで、HITAC 8400システムに2台の漢字ディスプレイを接続し、国会会議録情報などの校正作業を行なっている。

4.1.4 機械辞書

日本語情報の入力にはエラーがつきものであるから、ミス入力を削減する方法として1情報あたりの打鍵数を減少させることにより、エラーの発生件数を低下させることができる。そのために機械辞書という考え方が工夫された。これは五十音ソートのためのキーワードのふりがな、人名に対する所属機関名や役職名など主となるデータ項目から一義的に決定できるデータ項目、雑誌名や都道府県名のようにコード化可能なものはあらかじめ辞書ファイルを準備しておき、処理の段階で該当するデータ項目を抽出し、挿(そう)入したり置き換えたりする方法である。この方法は、辞書ファイルを常に最新情報にメンテナンスする必要があるが、入力情報量が少なく済み、エラーの発生件数を減少させる有効な手段である。

4.2 処理システム

一般に日本語を取り扱う内部処理は、かなや英数字だけを取り扱う処理と大きな相違はない。ファイルの更新や検索処理においても同様である。しかし漢字の配列基準が統一され

ていない現在、ソートは個々のシステムごとに目的に合うよう工夫している。

コンピュータに人間が日常使用していることばの意味や表現を人間が取り扱うのと同じように処理させ、コンピュータと人間との情報交換を容易にしようという要求は根強く、コンピュータユーティリティを開く重要な鍵(かぎ)となっている。通商産業省のパターン情報処理システム研究開発プロジェクトにおいても、パターン認識とともに、自然語情報処理を一つのテーマとしている。

この分野には次のようなものがある。

- (1) 表音文字から表意文字への言語変換
- (2) 機械翻訳
- (3) 自然語による情報検索、特に質問応答システム

4.2.1 自然語情報処理の問題点

自然語を取り扱うとき、自然語の持つあいまい性(Ambiguity)の問題は避けて通れない。

- (1) シンタックス上のあいまい性

一つの文(または文の構成単位)が二とおり以上の構文を持ちうること。

- (2) セマンティクス上のあいまい性

構文解析を経た文の意味が一意に保証されない。

- (3) 音韻上のあいまい性

同音異義語ばかりでなく、次のような構文認定上のあいまい性もあり、単位語認定誤りの原因となる。

カネオクレタノム → (金送れ、頼む),
(金を呉れた、飲む)

4.2.2 速記自動反訳システム

言語変換の具体例として速記自動反訳システムの構成と機能について述べる。

- (1) システムの概要

速記用タイプライタを打鍵して作成された速記記号列をコンピュータに入力し、自動的に漢字かな交じり文に反訳して速記録を作成するシステムである。

本システムの構成と処理方法は、図2に示すとおりである。

- (2) 速記記号の特質

- (a) 単音節を表わす「基本音節」と、2音節以上または特定の意味を持つ「略語」とから成り立っている。

- (b) 同一速記記号の訳が二とおり以上あるもの、訳の一部にかっこ書きを有するものなど、前後の文脈で正しい訳語を選択しなければならない。

- (c) 音を一部省略したり、くり返しを省略したり高速打鍵のための打ち方のきまりが多い。

- (3) 反訳の機能

連続して入力される表音記号(一部表意記号)で表わされた話しことばを構文解析により単位語の認定、品詞情報の認定を行なって、漢字かな交じり文章に変換している。本システムでは形態素分析法と呼ばれる手法を用いている。

反訳の機能として次の処理を行なう。

- (a) 漢字で書くべき語の漢字化
- (b) 外来語・外国語などのかたかな化
- (c) 数詞を表わすものの漢数詞化
- (d) 現代かなづかいの表記変換

〈ワ、オ、エ〉 → 〈は、を、へ〉

- (e) 前後の文脈に適した「かっこはずし」、多義性の選択

- (f) 文章を読みやすくするための読点の自動挿入

- (g) その他打ち方のきまりに対する処理

- (4) 形態素分析法

形態素分析法は次のものを用いて行なわれる。

- (a) 単位語辞書

見出し(速記記号の連系)とそれに対する品詞、訳などの情報を用意したもの。

- (b) 文法表

3型文法(有限状態文法)を遷移表の形式にしたもので、品詞数203、状態数96である。

解析の手法としてPDS(Push Down Storage)を利用しており、入力速記記号列に合致するものを単位語辞書から探索し、文法表を参照して先行語と後続語の品詞が続きうるかどうかが検定し、整合するものだけをPDSに積んでいく。こうして表意語から次の表意語まですべてPDSに積まれたとき、いちばん上にあるものが単位語の連続である。

- (5) 数詞処理

数詞類語彙(い)表と数詞構成規則テーブルにより形態素分析法と同様の処理を行なう。

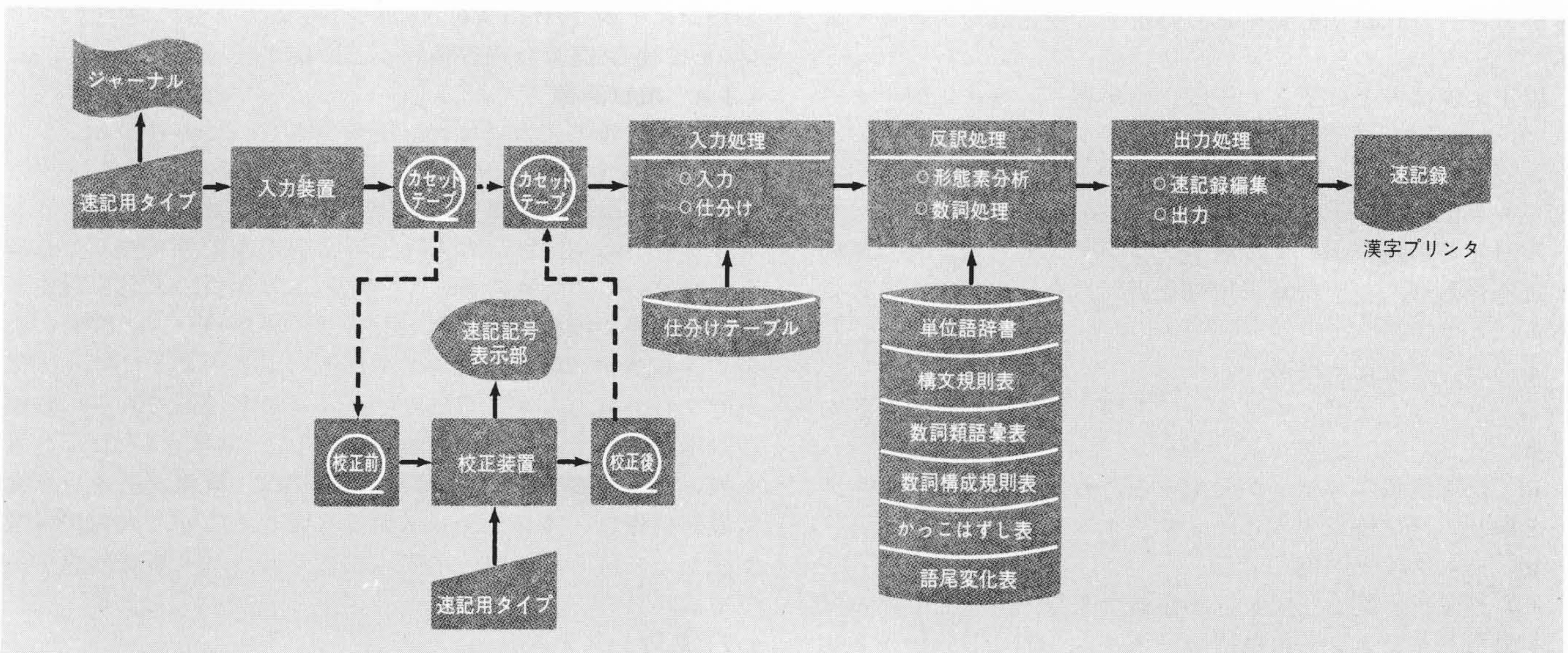


図2 速記自動反訳システムの構成と処理方式

各種辞書およびテーブル類の性能が、反訳精度を

決定づける大きな要因である。

Fig. 2 Configuration and Process Pattern of Stenograph to Japanese System

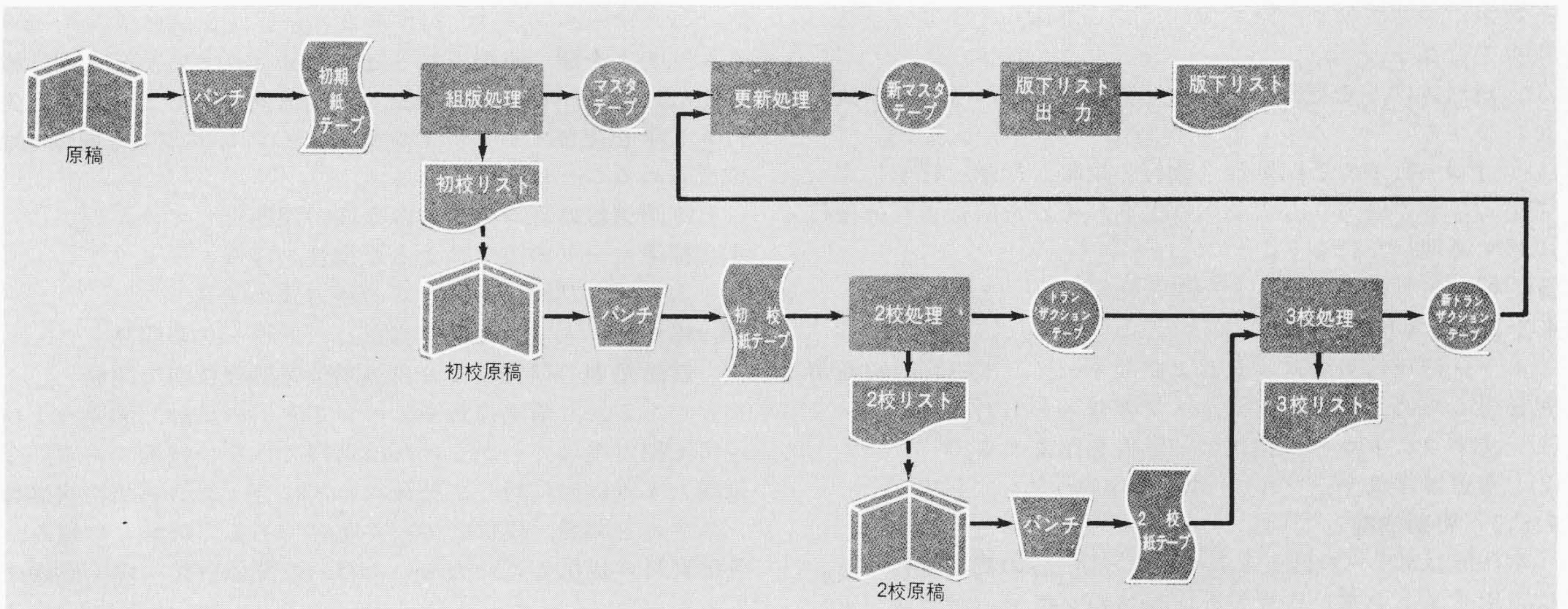


図3 日本語文ドキュメント自動組版システム 校正情報の完全化を図ってからマスター磁気テープの更新することにより、ターンアラウンドタイムの短縮を図っている。

Fig. 3 Typesetting Program System for Japanese Document

4.2.3 情報検索

情報検索では無限に近いことばやイメージ情報を整理・分類することがシステムの死命を制しており、システム建設においても索引作業に膨大な数の熟練した専門家を必要としている。また質疑応答による情報提供サービスが一般に受け入れられるためには、特別なコード化を要しない問合せ方式と、漢字かな交じり文や図、グラフなどのイメージ情報の出力が必要である。

- 情報検索システムに日本語を取り入れたときのメリットは
- (1) コード、同音異義語がなく理解度が高まる。
 - (2) 方式によっては検索処理速度、検索精度が向上する。

ここで(2)については、情報検索システムの適用する索引方式と密接な関係にあり、単に日本語の採用だけを軽々しく結論づける訳にはいかない。また問合せ方式において、日本語情報処理の入力システムの問題、自然語情報処理の問題が十分解決されていない段階において、日本語で問合せを行なうことは必ずしも得策とはいえない。情報検索システムに日本語を取り入れようという要求は、第1段階では検索結果の出力を漢字かな交じり文で表現することにより情報の理解度を高めようという方向に進むものと思われる。次段階では新入力方式の開発や自然語情報処理の問題解決とあいまって、日本語による問合せも可能となろう。

4.3 出力システム

日本語情報の出力システムは、日本語情報処理システムの中で最もコンピュータ化の効果の表われるところである。編集組版など事前にプログラム化されたルールに従って高速に大量データを処理し、版下作成を行なうシステムは植字工の不足などから急速に普及するものと思われる。また増大する各種ドキュメントをコンピュータで編集組版し、タイムリーな更新を行なうシステムは大量ドキュメントを扱う企業では必要不可欠となる。ここでは日立製作所のHDESに基づき編集組版システムの機能について述べる。

4.3.1 組版機能

- (1) 左横書きの組版を行ない、A4判を標準サイズとしてA5, A6, B5, B6の各判を選択する。
- (2) 文字の大きさは3種類、書体は漢字については明朝体、

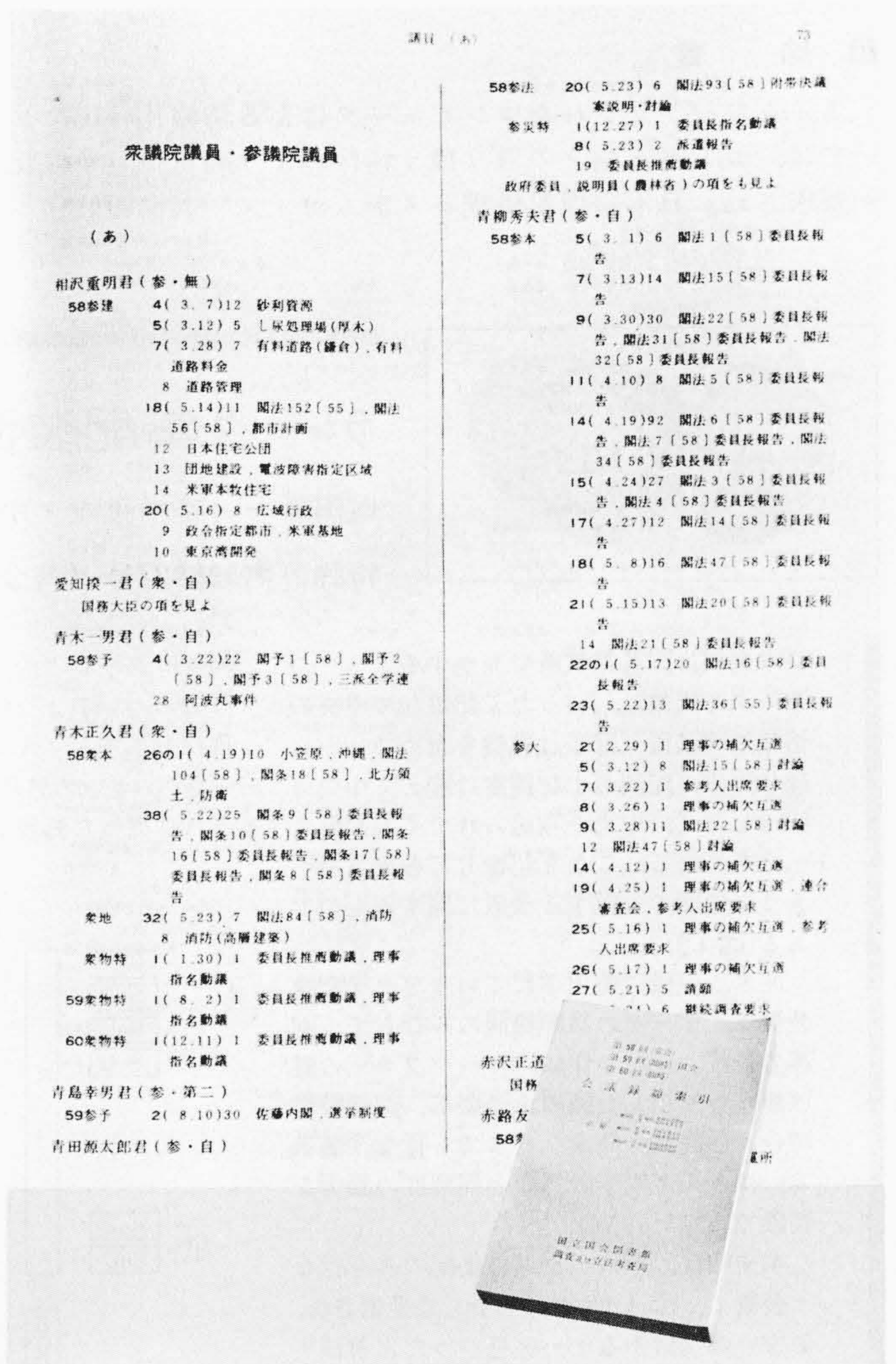


図4 機械編集された国会会議録総索引(議員名による索引の一部) 国立国会図書館におけるHITAC 8400システムで編集され、刊行された国会会議録総索引を示す。

Fig. 4 Computer Edited Directory of Diet Proceedings (for Indexing by each member's name)

