58－S003

文章情報データベース総合利用
調査研究報告書

昭和59年3月

JIPDEC

財団法人 日本情報処理開発協会

この報告書は，日本自転車振興会から競輪収益の一部である機械工業振興資金の補助を受けて昭和５８年度に実施した「文章情報データベースの総合利用に関する調査研究」の成果をとりまとめたものであります。

はじめに

近年，国際情勢の変化は，即時，わが国各界に大きく波及し，こうした事態への迅速・的確な対応はもとより，事前に兆候を把握できる体制づくりが急務となっている。

このためには，内外の情報資源を活用し，常時こうした動向を把握できる情報処理システム体制を確立する必要がある。

とくに，記事情報等の文章情報をデータベース化し，コンテンツ・アナリシス等の高度な手法を駆使して情報内容を分析し，客観的な判断材料を求めることは有効と思われる。

また，これらの情報処理過程においては，なるべく人手を介さない，自動的な機械処理が可能な形が望まれている。

こうしたことから，本事業においては，文章情報データベースを効果的に利用するに必要な総合解析システムを開発することを目的として，各種調査研究並びに機能開発等を実施している。

初年度は基礎的な調査研究を実施し，第２年度はこれに基づく課題を研究するとともに海外先進例の調査やデータベース作成・更新システムの研究開発を実施した。

第３年度として，本年度はひきつづき各種機能研究や海外調査を行うとともにデータベース検索システム，定量化分析モデルシステムの開発，文章情報の各種解析処理実験等を行った。

今後は，各種調査研究と並行してシステム機能の整備・拡大を図り，より効果的，実用的な総合解析システム構築に向けて研究開発を推進していくこととしたい。

最後に，本調査研究にあたって，ご指導，ご協力いただいた委員並びに関係各位に感謝の意を表します。

昭和５９年３月

「文章情報データベース総合利用調査委員会」委員名簿

（順不同，敬称略）

委員長	渕　一博	(財)新世代コンピュータ技術開発機構研究所長
委員	田中穂積	東京工業大学工学部情報工学科助教授
〃	矢田光治	知識情報研究所所長
〃	所沢仁	(財)工業開発研究所主任研究員
〃	小川芳樹	(財)日本エネルギー経済研究所研究部 第6研究室研究員
〃	小沢大二	国際協力事業団研修事業部次長
〃	木地三千子	日本経済研究センター特別研究員
〃	神尾達夫	日本経済新聞社データバンク局記事情報部次長
〃	知念利夫	日本経済新聞社データバンク局情報業務部
〃 前	小池康夫	(株)市況情報センター情報管理部長
〃 後	大倉健治	(株)市況情報センター開発第2部長
〃	海老原悦夫	日本貿易振興会企画部電子計算機室
〃	瀬谷重信	日本電信電話公社関東電気通信局データ通信部長
〃	山本恵美子	ソフトウェア研究会副会長
〃	木南公統司	(株)開発計算センターシステム管理部課長
〃	長谷川亨	(株)ソフトウェア・リサーチ・アソシエイツ 開発本部開発第5部長
〃	中瀬純夫	リソース・シェアリング(株)開発技術第2部長
〃	竹内憲	日本タイムシェア(株)第1システム事業部長付
〃	小泉秀雄	日本ハネウェル・インフォメーション・ システムズ(株)マーケティング推進部次長
〃	渡辺龍雄	通商産業大臣官房情報管理課長
〃	藤森聿子	通商産業大臣官房情報管理課 政策情報システム室長

委員	佐藤安夫	通商産業大臣官房情報管理課 政策情報システム室企画係長
〃	石川敬子	通商産業大臣官房情報管理課 政策情報システム室電子計算機専門職
〃	栗川正仁	通商産業大臣官房情報管理課計画班第2係長
〃	市川隆	(財)日本情報処理開発協会技術調査部長
〃	難波正之	(財)日本情報処理開発協会技術調査部次長

「文章情報データベース定量化利用研究専門委員会」委員名簿

（順不同，敬称略）

主査	山本毅雄	図書館情報大学図書館情報学部教授
委員	石川徹也	図書館情報大学図書館情報学部助教授
〃	神尾達夫	日本経済新聞社データバンク局記事情報部次長
〃	海老原悦夫	日本貿易振興会企画部電子計算機室
〃	木南公統司	㈱開発計算センターシステム管理部課長
〃	長谷川亨	㈱ソフトウェア・リサーチ・アソシエイツ 開発本部開発第5部長
〃	中瀬純夫	リソース・シェアリング㈱開発技術第2部長
〃	竹内憲	日本タイムシェア㈱第1システム事業部長付
〃	佐藤安夫	通商産業大臣官房情報管理課 政策情報システム室企画係長
〃	石川敬子	通商産業大臣官房情報管理課 政策情報システム室電子計算機専門職
〃	所沢仁	㈶工業開発研究所主任研究員
〃	湯浅俊昭	㈶日本エネルギー経済研究所第6研究室長
〃	小川芳樹	㈶日本エネルギー経済研究所第6研究室研究員
〃	難波正之	㈶日本情報処理開発協会技術調査部次長

「機械翻訳システム研究専門委員会」委員名簿

（順不同，敬称略）

委員	矢田光治	知識情報研究所所長
〃	田中穂積	東京工業大学工学部情報工学科助教授
〃	横山晶一	電子技術総合研究所パターン情報部推論システム研究室
〃	中瀬純夫	リソースシェアリング㈱開発技術第2部長
〃	山本恵美子	ソフトウェア研究会副会長
〃	知念利夫	日本経済新聞社データバンク局情報業務部
〃（前）	小池康夫	㈱市況情報センター情報管理部長
〃（後）	大倉健治	㈱市況情報センター開発第2部長
〃	瀬谷重信	日本電信電話公社関東電気通信局データ通信部長
〃	竹内憲	日本タイムシェア㈱第1システム事業部長付
〃	池田泰則	コンピュータサービス㈱システム機器販売事業部特機販売部次長
〃	木南公統司	㈱開発計算センターシステム管理部課長
〃	長谷川亨	㈱ソフトウェア・リサーチ・アソシエイツ開発本部開発第5部長
〃	佐藤安夫	通商産業大臣官房情報管理課政策情報システム室企画係長
〃	栗川正仁	通商産業大臣官房情報管理課計画班第2係長

目　　次

1. 事　業　概　要

1. 事 業 概 要

1.1 目的と背景

コンピュータ情報処理は，広く深く社会に浸透し，情報化社会は名実ともに確立されつつある。

近年のオフィス・オートメーションやパーソナル・コンピュータの急激な普及にもみられるように，職場や家庭における一般利用者の比較的容易な操作によるコンピュータ利活用が可能な時代となった。

こうした状況の背景には，ハードウェア，ソフトウェアの技術向上に伴う諸条件整備はもとより，情報資源に対する社会及び個人の関心，意識の向上が大きく作用していると思われる。

情報そのものの重要性が認識され，情報量が拡大されるに伴い，データベースや，情報検索システムといった情報流通機能の需要も拡大するのは必須であり，益々増大する情報利用者のニーズに対応すべく，情報の量的，質的整備と簡便な利用体制づくりを中心とした情報処理システムの確立が急務となってきた。

とくに文章情報データベースは日本語情報処理の普及や海外からの文献情報データベース等の導入により，その蓄積及び利用は急速に高まるものと思われる。

それらを効果的に利用するには通常の情報検索に加えて，データベースの持つ各種の情報をコンテンツ・アナリシス等の高度な分析（キーワードの頻度の時系列分析，出現頻度の相関分析，意味論的分析）をし，利用することが極めて重要である。

このため，文章情報データベースを効果的に利用するに必要な総合解析システムを開発することを目的として，調査研究を実施する。

1.2 実施経過

本事業の計画概要は図１－１に示すとおりであり，これまでの実施経過を以下に示す。

(1) 昭和５６年度

「文章情報データベース総合利用調査委員会」を設置して，本調査研究の基本計画を策定し，事業の推進とりまとめを行った。

また，委員会メンバーを中心とするワーキング・グループで以下のテーマに基づいて調査研究を実施した。

① カントリーリスク，エネルギー動向把握のための文章情報コンテンツ・アナリシス手法の研究

- キーワード自動抽出法の研究
- 入力データ作成上の課題
- 既存システム利用事例研究
- シソーラス辞書作成の研究
- 定量化利用方法論の研究
- 新記事情報利用システムの研究

② 海外情報の有効活用，問題別把握，分析を行うための翻訳システム実用化の基礎研究

- 実用可能性の研究
- 機械翻訳技術の研究
- 構文解析技術の研究
- 入出力インターフェイスの研究
- ＬＣ－ＭＡＲＣの実験準備

(2) 昭和５７年度

前年度と同様「文章情報データベース総合利用調査委員会」により，全体の推進，統括を行った。

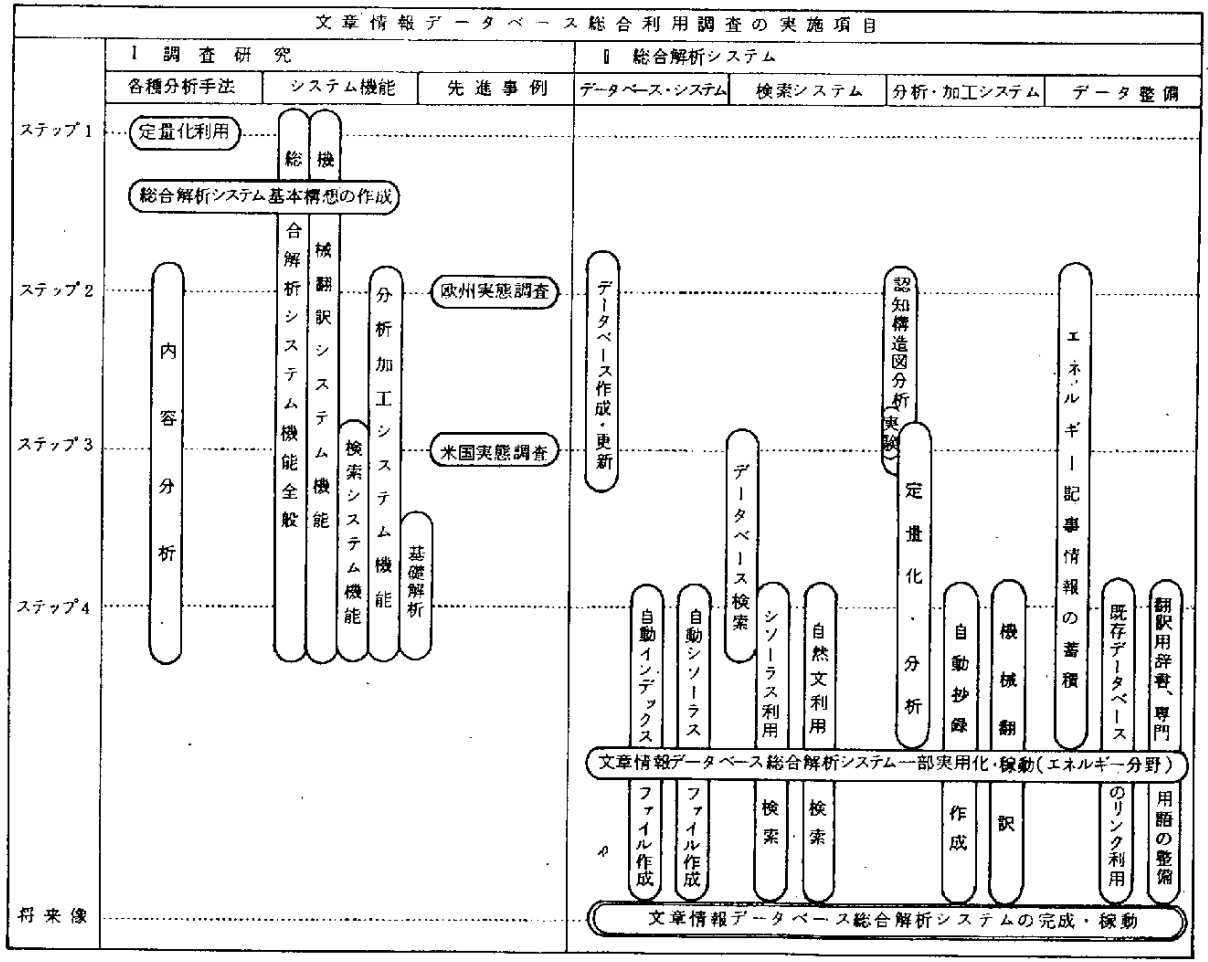

文章情報データベース総合利用調査の実施項目
Ⅰ 調査研究
Ⅱ 総合解析システム
各種分析手法
システム機能
先進事例
データベース・システム
検索システム
分析・加工システム
データ整備
ステップ1
ステップ2
ステップ3
ステップ4
将来像
定量化利用
総
機
総合解析システム基本構想の作成
内容分析
合解析システム機能全般
械翻訳システム機能
検索システム機能
分析加工システム機能
基礎解析
欧州実態調査
米国実態調査
データベース作成・更新
認知構造図分析実験
エネルギー記事情報の蓄積
自動インデックス
自動シソーラス
データベース検索
シソーラス利用
自然文利用
定量化・分析
自動抄録
機械翻
既存データベース
翻訳用辞書、専門
文章情報データベース総合解析システム一部実用化・稼動(エネルギー分野)
ファイル作成
ファイル作成
検索
検索
作成
訳
のリンク利用
用語の整備
文章情報データベース総合解析システムの完成・稼動

図1－1　計画概要

委員会の下に「文章情報データベース定量化利用研究専門委員会」並びに「機械翻訳システム研究専門委員会」の２つの専門委員会を設置して，研究を実施した。

「文章情報データベース定量化利用研究専門委員会」では，

- 国際紛争データと危機管理システムの研究
- General Inquirer におけるデータ作成と解析
- 認知構造図手法の研究と適用実験

について事例研究を行うとともに，総合解析システムのうちデータベース作成・更新システムの開発を実施した。

つまり近年広く各界の注目を集めているエネルギー，カントリーリスク関連分野をモデルに記事情報の整備及びデータベースの構築を行った。

「機械翻訳システム研究専門委員会」では，

- 辞書構造の研究と作成上の課題
- 〝翻訳システム先進事例の研究
- 構文解析法と辞書の役割

について研究を行った。

また，①文章情報の有効利用法と機械翻訳のアルゴリズム，②文章情報解析システムにおける構文解析と用語データベースの役割をテーマに，欧州を中心とする先進事例の調査を行い，今後の参考に資した。

(3) 昭和５８年度

５６，５７年度に引き続き委員会並びに２つの専門委員会を設置して調査研究を推進した。

「文章情報データベース定量化利用研究専門委員会」においては，

- 記事情報インデックスのキーワード出現頻度数分析
- キーワード自動抽出システムの現状
- 文章情報内に出現する自然語の頻度分析

について事例研究を行うとともに，５７年度開発したデータベース作成，更新システムに付加する検索システム並びに定量化分析のモデルシステムの開発を実施した。また，エネルギー記事情報の本格整備を図り，データベースを構築した。

「機械翻訳システム研究専門委員会」においては，翻訳アルゴリズムや文法規則，辞書等の研究を次の実験を通して行った。

- ○　ＰＲＯＬＯＧを利用した文章解析処理の実験
- ○　エネルギー分野における用語収集実験

また，米国の研究機関を中心に自然言語処理技術の動向やデータベース作成処理の実態調査を行い参考に資した。

なお，５８年度本調査研究の経過概要は表１－１に示す通りである。

表１－１　昭和５８年度調査実施経過表

項目＼月	58/4	5	6	7	8	9	10	11	12	59/1	2	3
Ⅰ　実施計画の策定	←	→										
Ⅱ　委員会の開催												
• 文章情報データベース総合利用調査委員会		19○										28○
• 文章情報データベース定量化利用研究専門委員会				6○							7○	22○
• 機械翻訳システム研究専門委員会								16○				8○ 21○
Ⅲ　文章情報総合利用の研究												
• キーワード自動抽出システムの現状							←		→			
• 記事情報インデックスのキーワード出現頻度数分析							←			→		
• 文章情報内に出現する自然語の頻度分析							←		→			
• ＰＲＯＬＯＧを利用した文章解析処理の実験				←								→
• エネルギー分野における用語収集実験				←								→
Ⅳ　海外調査員の派遣							← →					
Ⅴ　総合解析システムの開発												
• データベース検索システムの開発				←						→		
• データベース定量化分析システムの開発					←						→	
• エネルギー記事情報の整備				←						→		
Ⅵ　報告書の作成										←		→

1.3 本報告書の構成

本報告書は次の4章からなっている。

まず第1章で事業概要をのべ，第2章では基本構想に基づく総合解析システムの開発手順と本年度開発したデータベース検索システム並びに定量化分析システムの概要と，エネルギー記事情報のデータベース作成についてふれた。

第3章は，文章情報データベースの総合的な利用に資するため，文章情報，特にキーワードを中心とした頻度分析の研究，自動抽出システムの現状調査，文の解析処理実験の結果，および米国における調査成果等を，各テーマ毎にまとめた。

第4章は，これまでの成果を踏まえて問題点を抽出し，現状認識に基づく今後の方向性を考察して総合解析システム開発への具体的アプローチをまとめた。

2. 文章情報データベース総合解析システムの構築

2. 文章情報データベース総合解析システムの構築

2.1 総合解析システムの開発手順

文章情報総合解析システムは，大量の文章情報をデータベースに蓄積し，これを検索・内容分析することにより事態の客観的な把握，将来動向の予測，仮説の検証等に役立つデータを提供することを目的とし，政策の立案や決定，企業の経営計画の策定等に際して有力なサポートシステムとして機能することをめざしている。

文章情報の総合解析は，概念的には文章を構文的・意味論的に解析するプロセス（基礎解析）と，基礎解析の結果を分析・編集・合成して利用者が求める情報を希望する形態及び言語で出力するプロセス，の2段階の過程を経て達成される。後段のプロセスは，検索・内容分析と翻訳との2つのプロセスに大別され，利用者の選択によって必要な処理が行われる。また，いずれのプロセスにおいても辞書や文法規則等の知識ベースを必要とする。

従って，基礎解析，検索，内容分析，翻訳，知識ベースの5つが総合解析システムとして具備すべき機能である。（図2.1－1参照）

このような基本構想の基で，本調査研究では各種分析手法，システム機能等について海外も含めた形で先進事例調査を行い，分析手法については実データを使用して一部実験を行ってきた。また，調査結果及び現在の技術水準等を踏まえ，可能なところからシステム開発を実施してきた。これまでに開発したシステムは以下のように位置付けられる。

(1) データベース作成・更新・修正サブシステム

本サブシステムで取り扱うデータは海外のエネルギー関連情報を基に，専門家により日本語で作成したインデックスである。

従って，データベースからみると総合解析システム概念図の二次情報データベースに該当するものである。

(2) データベース検索システム

本システムは概念図の検索部分のベースになるものであり，システムの詳細は後述する。

(3) 定量化分析システム

本システムは内容分析の基礎的な機能として，キーワードを基にした頻度の時系列・クロスセクション分析を可能とするものであり，さらに必要に応じ主成分分析，回帰分析をも可能とするものである。

本章では本年度開発した上記(2)及び(3)のシステムとデータ整備について述べることとする。

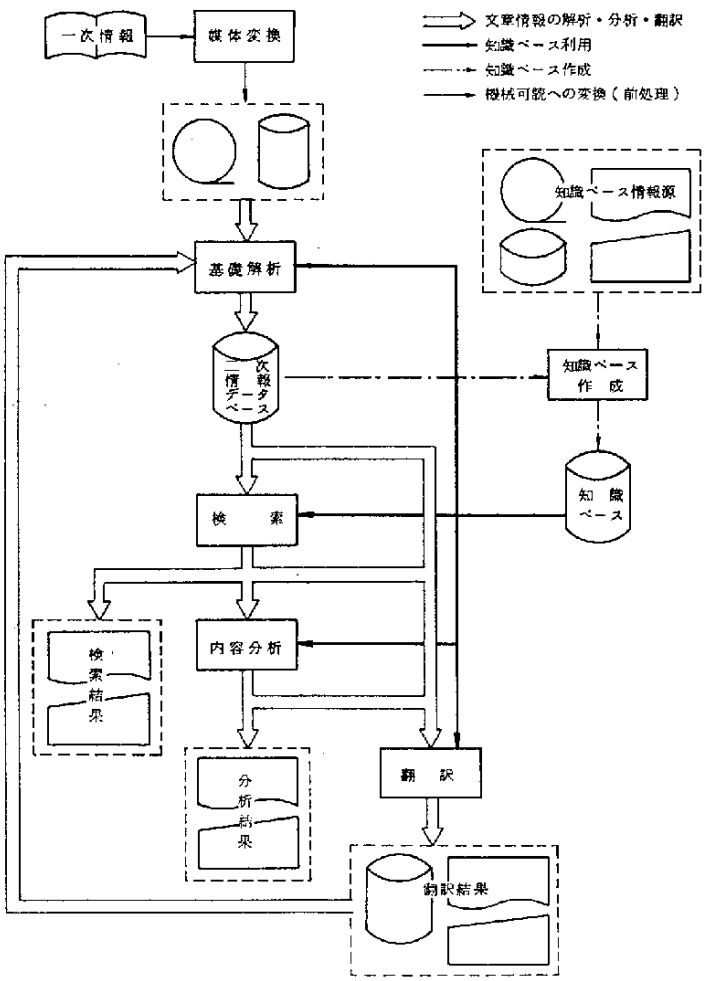


図 2.1－1　総合解析システム概念図

2.2　データベース検索システムの開発

2.2.1　システムの目的

文章情報（記事情報）は内容豊かな情報源であり，統計数値情報よりも速報性を持っている。例えば，エネルギー問題，経済問題，カントリー・リスク問題などの分野を考えた場合，網の目のようにはりめぐらした情報網から，これは平常な流れとは異った動きであるということを敏速にとらえる仕組みを作り上げることはきわめて重要と考えられる。このような動向をキャッチさえできれば，それを出発点としていろいろな手段を講じて調べ上げることはそれほど困難なことではない。

しかし，文章情報の場合，広範な情報が含まれているので，情報密度の点ではかなり薄いということができよう。従って，上述のような意味での文章情報の有効利用を考えた場合，一般検索による利用もさることながら，特定のテーマに基づき文章情報を集約化して利用することが重要である。その方法として文章情報から様々な情報をとり出して定量化し，さらに多変量解析等を適用していく方向が考えられる。

このため，モデルとしてエネルギー分野を取り上げ，５７年度開発したデータベース作成・更新システムを活用して，引き続きデータベース検索システムの開発を行った。

2.2.2　情報の範囲

情報の収集は海外で発行されている主要５０紙誌の中からエネルギー問題を論及した主要記事を対象としている。

情報の内容は，上記主要記事から一定のルールに従ってコーディングしたものであり，大別すると次の３種類となる。

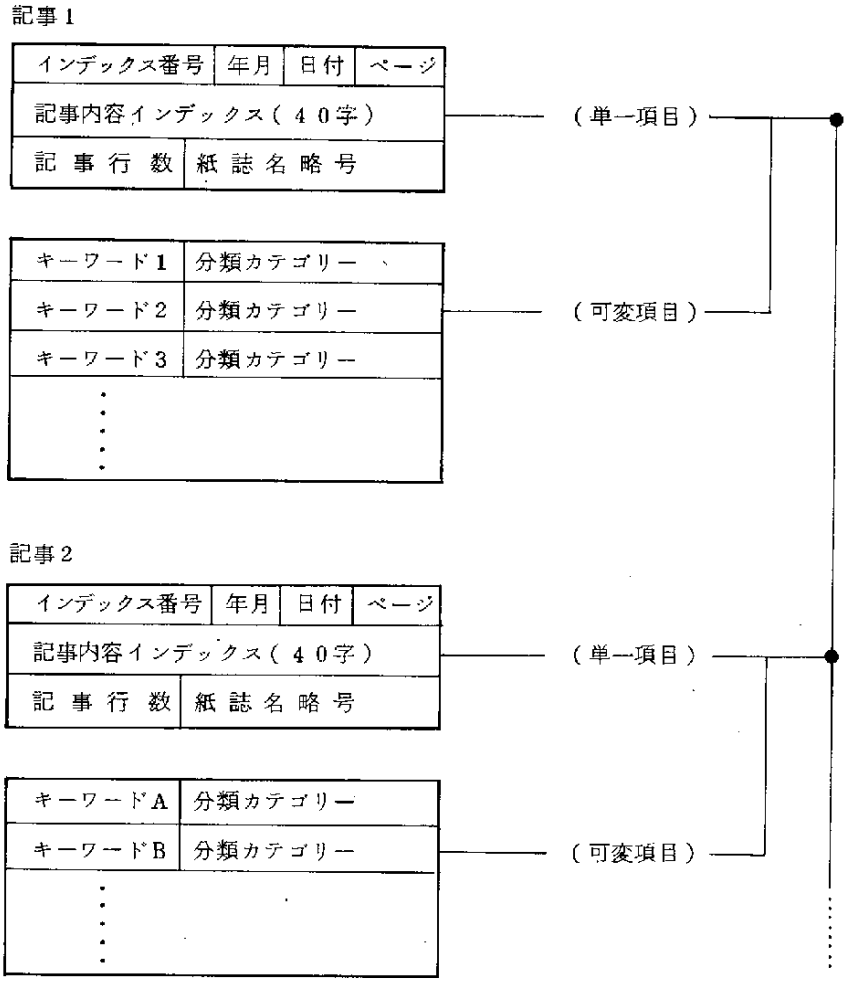
①　データの所在に関する情報

②　データの具体的内容に関する情報

③　データを入手し利用するための情報

なお，情報の項目及びその概要の一覧は表 2.2－1に示す通りである。

表 2.2－1　情報の項目一覧表



2.2.3　システムの機能

本データベース検索システムの機能としては，エネルギー記事情報インデックスの１次検索機能，論理検索機能，２次検索機能，頻度検索機能の４つがある。

検索システムの処理形態は，応答形式のインタラクティブな処理とする。

各サブシステムとその機能については図2.2－1に，各サブシステム間の情報関連図は図2.2－2に示すとおりである。

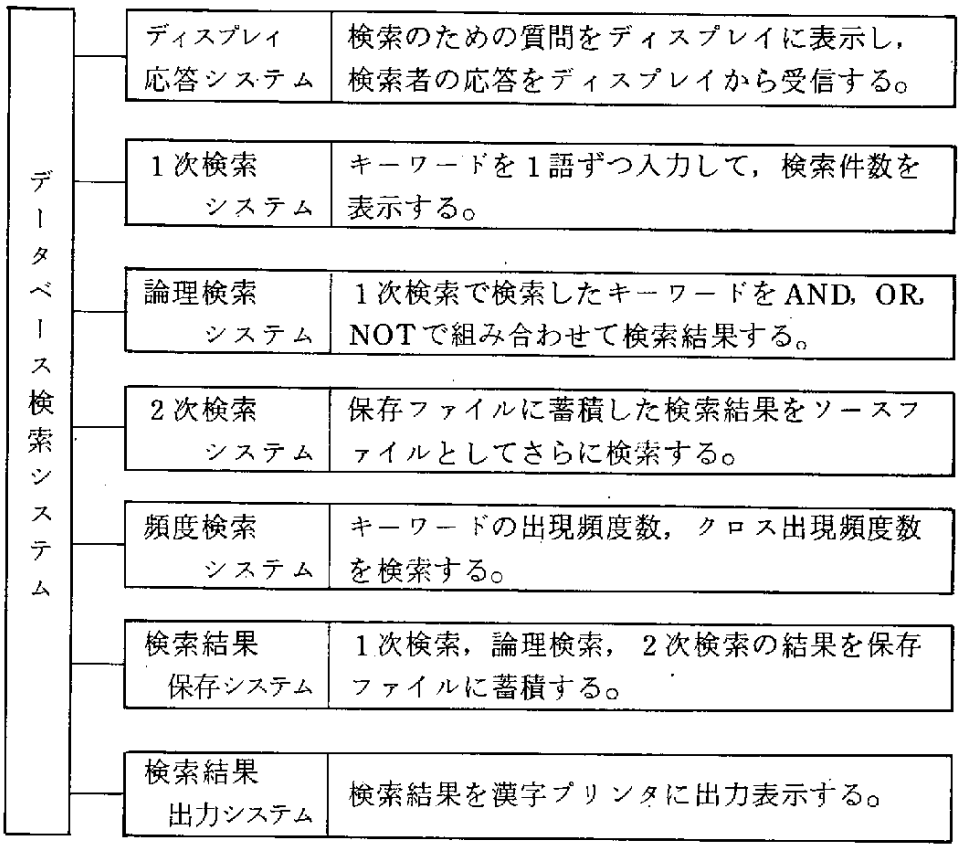


図2.2－1　データベース検索システム

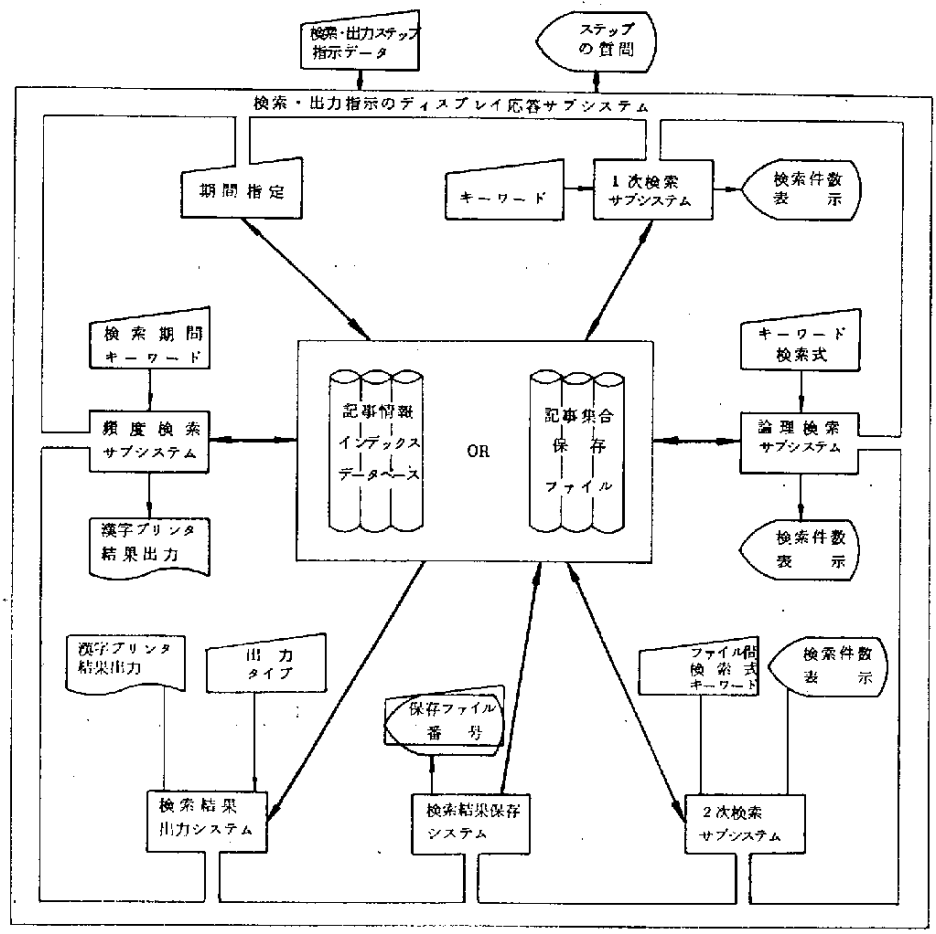


図 2.2－2　サブシステム間の情報関連図

2.2.4　システムの概要

各サブシステム単位に，サブシステムの機能，入力情報，出力情報，作成ファイル等の概要について述べる。

データベース検索システムの入出力情報の一覧は図 2.2－3 に示すとおりである。

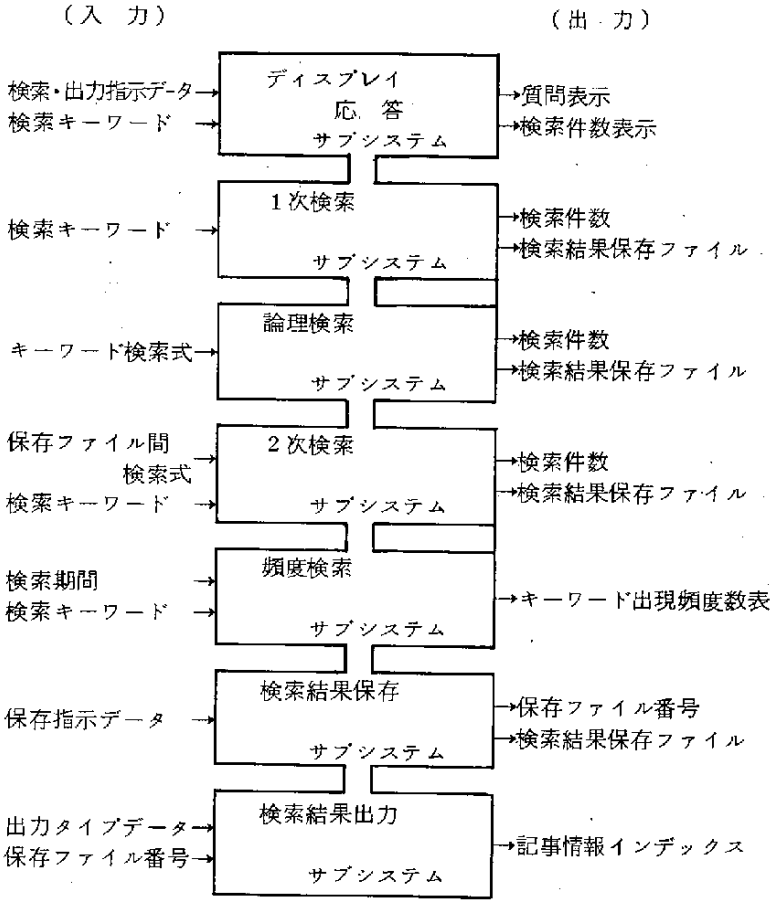
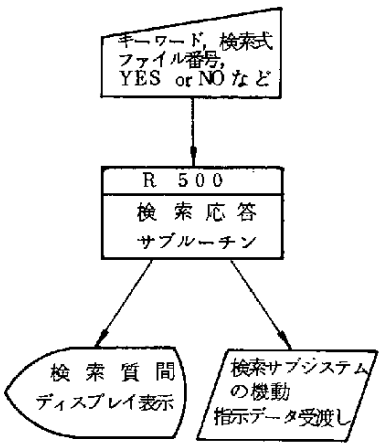


図 2.2－3　データベース検索システムの入出力

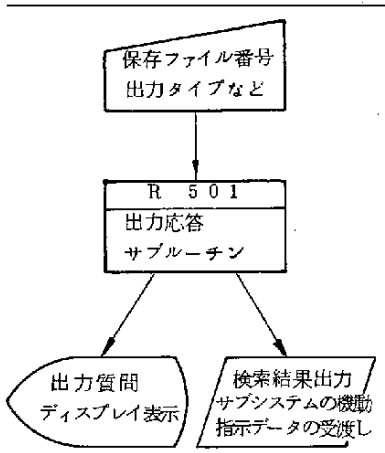
(1) ディスプレイ応答サブシステム

(a) ディスプレイ応答サブシステムの機能

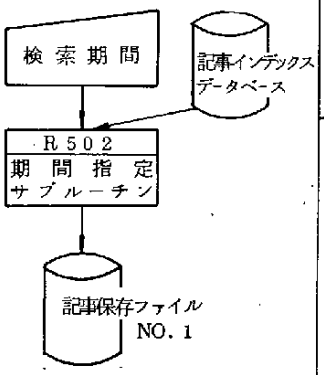
① 検索質問応答機能

処理フロー	ブロック名	機能
		当サブルーチンでは検索質問を表示し，ユーザの応答を受け，対応する検索処理を機動する。
	R 5 0 0	検索ステップを選ばせる質問を表示し，ユーザの応答に応じて各サブシステムを機動する。各サブシステムの検索段階に応じて質問を表示，キーワード，検索式，保存ファイル番号などユーザの応答に従って対応する処理を機動する。

② 出力質問応答機能

処理フロー	ブロック名	機能
		当サブルーチンでは，出力タイプ保存ファイル番号などを質問し，結果出力サブルーチンを機動する。
	R 5 0 1	保存ファイルに蓄積した検索結果を出力するため，保存ファイル番号と出力タイプを質問し，検索結果出力サブシステムを機動する。

③　期間指定機能

処理フロー	ブロック名	機能
		当サブルーチンでは，システム開始直後にデータベースの検索期間の絞り込みを行う。
	R502	記事インデックスの登録期間よりもせばめて，すべての検索を行うため該当する検索期間の記事インデックスを№1の保存ファイルに蓄積し，以降ソースデータ・ファイルを№1の保存ファイルとする。

(b)　ディスプレイ応答サブシステムの出力情報

ディスプレイ応答サブシステムでは，システム・ユーザと対話型で検索あるいは結果出力の指示データのやりとりを行う。ユーザの入力の負担を避けるため，検索や結果出力の諸段階に応じて動作の内容を応答形式とともに質問の形でディスプレイに表示している。以下に質問文の例を示す。

＜質問文の例＞

・期間指定を行いますか？

（入力形式：はい＝1，いいえ＝0）

・期間を指定して下さい。

（入力形式：7901－7903）

・行いたい検索を指定して下さい。

（入力形式：一次検索＝1，二次検索＝2，頻度検索＝3，

結果出力＝4）

・論理検索に使用するキーワードを入力して下さい。

・この記事集合を保存ファイルへ登録しますか？

（入力形式：はい＝1，いいえ＝0）

(c) ディスプレイ応答サブシステムの入力情報

(b)で述べたようにディスプレイ応答サブシステムでは，検索・出力動作の諸段階が質問の形で表示されるので，システム・ユーザは表示された入力形式に従って，ディスプレイから0，1など簡単な数字入力で応答を送信していけばよい。

検索のためのキーワードだけは，漢字キーワードで入力する必要がある。このためには，タブレット方式の漢字入力を持ったディスプレイか，かな漢字変換機能を持ったディスプレイを使用しなければならない。

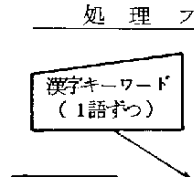
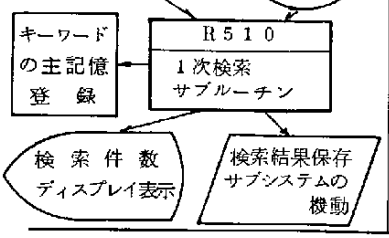
表示した入力形式に従ってシステム・ユーザからの応答を受信すると，応答に従って対応する処理ルーチンが機動する。

(d) ディスプレイ応答サブシステムで作成するファイル

期間指定を行った場合だけ，対象検索期間の記事インデックスをソース・データベースから検索して記事集合保存ファイルを作成する。期間指定の場合は，以後の検索でこの保存ファイルがソースデータベースとなる。

(2) 1次検索サブシステム

(a) 1次検索サブシステムの機能

処理フロー	ブロック名	機能
		論理検索に使用するキーワードを1語ずつ受信して該当記事インデックスの検索件数を表示する。
	R510	ディスプレイ応答システムから検索するキーワードを受取り，データベース検索質問文を作成して検索する。該当記事がある場合，検索キーワードを論理検索のため主記憶に保存し，検索結果を保存するかどうかを質問する。1次検索終了の指示を受信するまでは同じ動作を繰り返す。

(b) 1次検索サブシステムの入力情報

当サブシステムの入力情報は，ディスプレイ応答システムで受け取ったシステム・ユーザの応答が受け渡されてくる（以下サブシステムでも同様）。

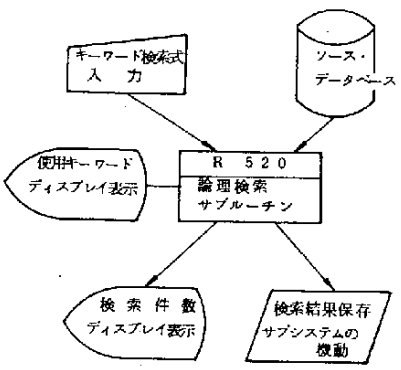
入力情報の内容は，1次検索のための漢字キーワード（1語），検索結果保存指示データ，1次検索終了指示データである。

(c) 1次検索サブシステムの出力情報

当サブシステムの出力情報は，キーワード1語検索を行って得られる該当記事インデックスの件数である。1次検索は，次に説明する論理検索に使用するキーワードを蓄積するため，該当インデックスをもつキーワードの確認を行うのがシステムの主目的である。

(3) 論理検索サブシステム

(a) 論理検索サブシステムの機能

処理フロー	ブロック名	機能
		1次検索で蓄積キーワードを組み合わせてキーワード検索式を作り，データベース検索を行う。
	R520	ディスプレイへ1次検索で蓄積したキーワードを表示し，キーワード番号と演算子を用いたキーワード検索式を受取る。式に基づいて検索文を作りデータベース検索を行う。該当記事インデックスが存在する場合，検索結果を保存するかどうか質問する。論理検索終了の指示を受信するまでは同じ動作を繰り返す。

(b) 論理検索サブシステムの出力情報

論理検索サブシステムでは，検索を行うため，1次検索で蓄積しておいた漢字キーワードをキーワード番号を付してディスプレイに

表示する。

論理検索を行うと該当記事インデックスの検索件数も出力する。

(c) 論理検索サブシステムの入力情報

論理検索サブシステムの入力情報は，キーワード検索式，検索結果保存指示データ，論理検索終了データである。

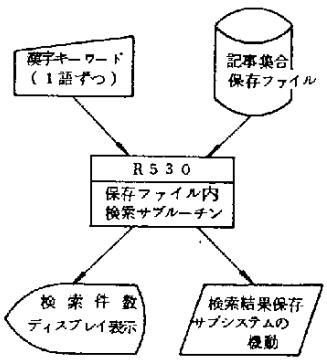
上述のようにディスプレイには1次検索で蓄積した漢字キーワードがキーワード番号とともに表示される。キーワード検索式は，キーワード番号と演算子“＋”（ＯＲ），“＊”（ＡＮＤ），“－”（ＮＯＴ），“（“，”）”を組み合わせた形で入力される。

例　（１＋２）＊３，（－１）＊２など

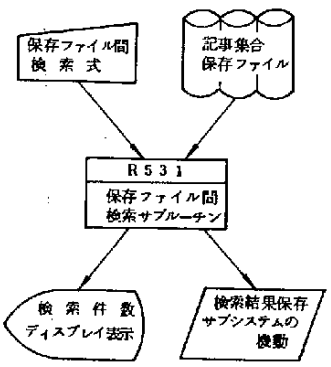
(4) 2次検索サブシステム

(a) 2次検索サブシステムの機能

① 保存ファイル内検索機能

処理フロー	ブロック名	機能
		1次検索，論理検索あるいは2次検索の結果を保存した記事集合保存ファイルをソースとして，キーワード1語ずつの検索を行う。
	R530	記事集合保存ファイルの番号をディスプレイから受けて，記事集合保存ファイルをオープンする。後は，1次検索と同じように検索キーワードをディスプレイから受け取り検索を行い，検索件数を表示する。該当記事インデックスがある場合は，検索結果を保存するかどうかを質問する。保存ファイルを変更して検索することも可能である。保存ファイル内検索終了データを受信するまでは検索を続ける。

② 保存ファイル間検索機能

処理フロー	ブロック名	機能
		1次検索，論理検索あるいは2次検索の結果を保存した記事集合保存ファイル間で，AND，ORの検索を行う。
	R531	ディスプレイから，保存ファイル番号と演算子を用いた保存ファイル間検索式のデータを受け取り検索を行う。検索件数を表示し，該当記事インデックスがある場合は，検索結果を保存するかどうかを質問する。保存ファイル間検索終了データを受信するまでは検索を続ける。

(b) 2次検索サブシステムの入力情報

当サブシステムの入力情報は，保存ファイル内検索のための保存ファイル番号（02～32），漢字キーワード（1語），保存ファイル間検索のための保存ファイル間検索式，及び検索結果保存指示データ，検索終了データなどである。

保存ファイル間検索式は，保存ファイル番号（02～32）と演算子"＋"（OR），"＊"（AND）の組み合わせで与えられる。

例 01＋02，05＊10など。

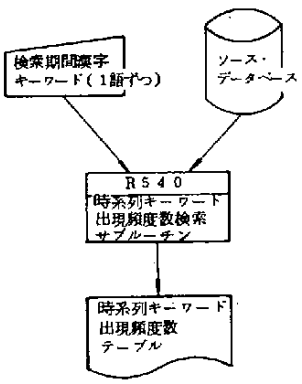
(c) 2次検索サブシステムの出力情報

当サブシステムの出力情報は，検索の結果与えられる該当記事インデックスの件数である。

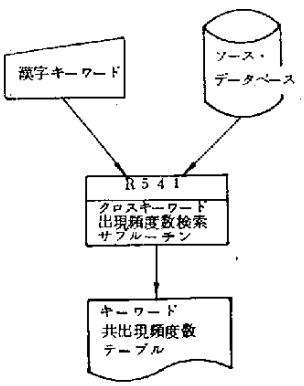
(5) 頻度検索サブシステム

(a) 頻度検索サブシステムの機能

① 時系列キーワード出現頻度数検索の機能

処　理　フ　ロ　ー	ブロック名	機　　能
		キーワードの出現頻度数を時系列で指定した期間内に関して検索し，出力する。
	R540	ディスプレイから検索期間を受信した後，漢字キーワードを1語ずつ入力し，指定された期間におけるキーワードの出現頻度数を時系列で検索する。検索結果は順次漢字プリンタに出力し，横に期間を，縦にキーワードを取った出現頻度数の表を出力する。検索終了データを受信するまで検索を続ける。

② クロスキーワード出現頻度数検索の機能

処　理　フ　ロ　ー	ブロック名	機　　能
		キーワードの共出現頻度数を指定した期間内に関して検索し出力する。
	R541	ディスプレイから，共出現を取るキーワード群を入力（横軸）する。以後はキーワードを1語ずつ入力し，指定された期間における横軸キーワード群との共出現頻度数を検索する。検索結果は，横軸キーワード群を出力して，出現頻度数を順次漢字プリンタに出力する。検索終了データを受信するまで検索を続ける。

(b) 頻度検索サブシステムの入力情報

当サブシステムの入力情報は，検索期間，漢字キーワード，検索終了データなどである。

(c) 頻度検索サブシステムの出力情報

当サブシステムの出力情報は，漢字プリンタへの時系列キーワ

ード出現頻度数テーブルとキーワード共出現頻度数テーブルである。

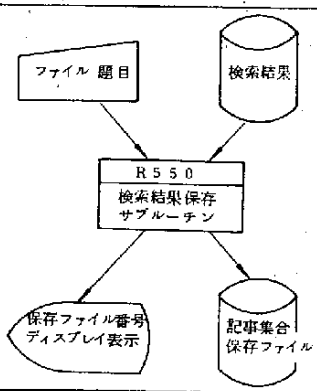
時系列キーワード出現頻度数テーブルは，横に期間を取り，縦に順次検索したキーワードを列記して，出現頻度数を出力する。

キーワード共出現頻度数テーブルは，横にキーワード群を並べ，縦に1語ずつキーワードを列記しながら横に並べたキーワードとの共出現頻度数を出力する。

出力例は2.4にまとめて表示する。

(6) 検索結果保存サブシステム

(a) 検索結果保存サブシステムの機能

処理フロー	ブロック名	機能
		1次検索，論理検索，2次検索での検索結果を記事集合ファイルとして保存する。
	R550	検索結果保存指示のデータを受信すると，ファイル内容を示す題目（15字以内）を入力させ，ファイルへ保存する。保存終了後，ファイル題目を主記憶に保存し，保存したファイル番号をディスプレイに表示する。

(b) 検索結果保存サブシステムの入力情報

当システムの入力情報は，保存したファイルの内容が後になってもわかるようにするため，15字以内の漢字文で与えられるファイル題目である。

(c) 検索結果保存サブシステムの出力情報

当システムの出力情報は，保存したファイルのファイル番号である。

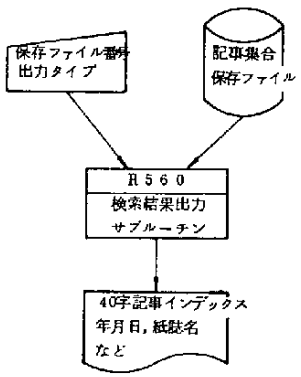
(d) 検索結果保存サブシステムで作成するファイル

ファイル番号01の記号集合ファイルは，期間指定を行った時のソースデータファイルとして利用されるので，02以降システムが

許容する範囲内の記事集合ファイルが検索結果保存指示に応じて作成される。

(7) 検索結果出力サブシステム

(a) 検索結果出力サブシステムの機能

処理フロー	ブロック名	機能
		検索結果を保存した記事集合ファイルから出力様式に従って検索結果を漢字プリンタに出力する。
	R560	ディスプレイから保存ファイル番号，出力タイプのデータを受信すると，出力様式に従って漢字プリンタへ記事インデックスなどの検索情報を出力する。結果出力終了のデータを受け取るまではこの動作を続ける。

(b) 検索結果出力サブシステムの入力情報

当サブシステムの入力情報は，結果出力をしたい保存ファイルの番号出力タイプ，結果出力終了データである。

(c) 検索結果出力サブシステムの出力情報

当サブシステムでは，１次検索，論理検索，２次検索を行って，保存した記事集合保存ファイルから４０字記事インデックス，掲載年月日，ページ紙誌名などを漢字プリンタに出力する。

出力タイプは次の２種がある。

＜タイプ１＞

- ○インデックス番号
- ○年月日
- ○ページあるいは紙面
- ○４０字記事内容インデックス
- ○記事行数
- ○紙誌名略号

＜タイプ２＞

○タイプ１の内容

○各種キーワード

2.2.5　コンピュータの機器構成

当システムで必要な機器構成を図２.２－４に示す。

検索システムのためには，中央処理装置，磁気ディスク装置（コンピュータシステムが使用するもの以外にデータベース用として１台），漢字プリンタ装置，漢字入出力の可能なディスプレイ・ターミナル及びコンピュータシステムの制御・監視のためのオペレータ・ステーションが構成機器である。

また，データベースの作成・更新をバッチ処理で行うためには，カード読取装置，磁気テープ装置（２台）の設置が必要である。

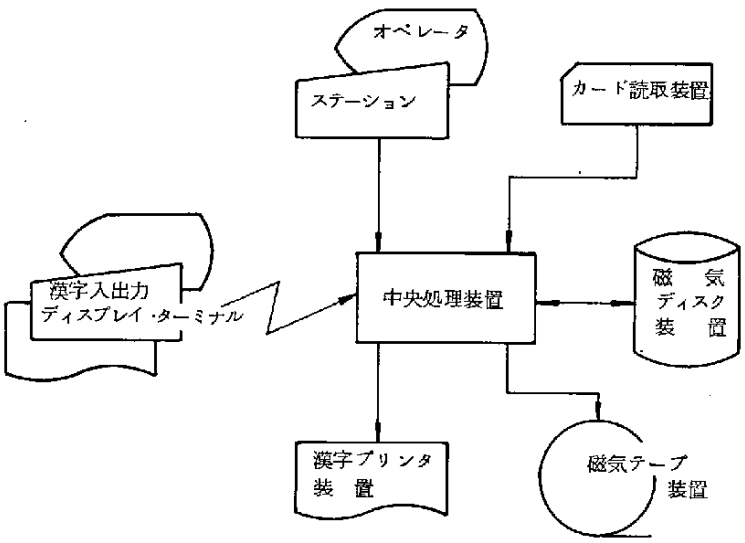


図２.２－４　コンピュータの機器構成

2.3　定量化分析システムの開発

2.3.1　システムの目的

文章情報総合解析システムとして具備すべき機能には大別して，基礎解析，検索，内容分析，翻訳，知識ベースの５つがある。

このうち，内容分析につき，１手法として情報内容の定量化による分析をとりあげ，機能研究，分析実験（3.1で示す）の結果を検討して定量化分析システムを開発した。

2.3.2　情報の範囲

定量化分析システムで使用する情報の範囲は，記事情報インデックスの頻度検索で得られるキーワードの出現頻度数である。データベース検索システムで整備した頻度検索で得られるキーワードの出現頻度数だけでは定量化分析のためのベースデータとして不充分であるため，頻度検索サブシステムの検索機能を拡充し，また生のキーワード出現頻度数に加工を加えて基礎情報として使えるようにした。

2.3.3　システムの機能

本定量化分析システムの機能としては，データ準備機能としてキーワード頻度数検索機能，キーワード頻度数保存機能，キーワード頻度数データ加工機能，生データ及び加工データグラフ出力機能があり，分析機能として多変量解析手法の中から主成分分析機能，回帰分析機能を整備する。

定量化分析システムの処理形態は，応答形式のインタラクティブな処理とする。

各サブシステムとその機能については図2.3－1に，各サブシステム間の情報関連図は図2.3－2に示すとおりである。

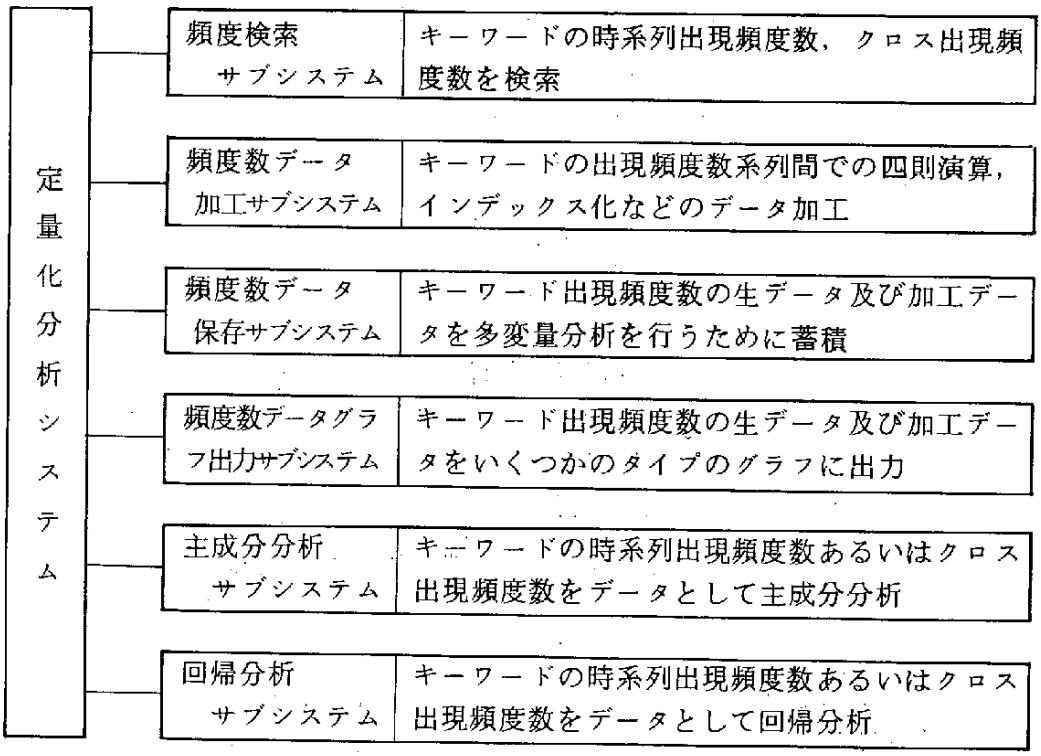


図 2.3－1　定量化分析システム

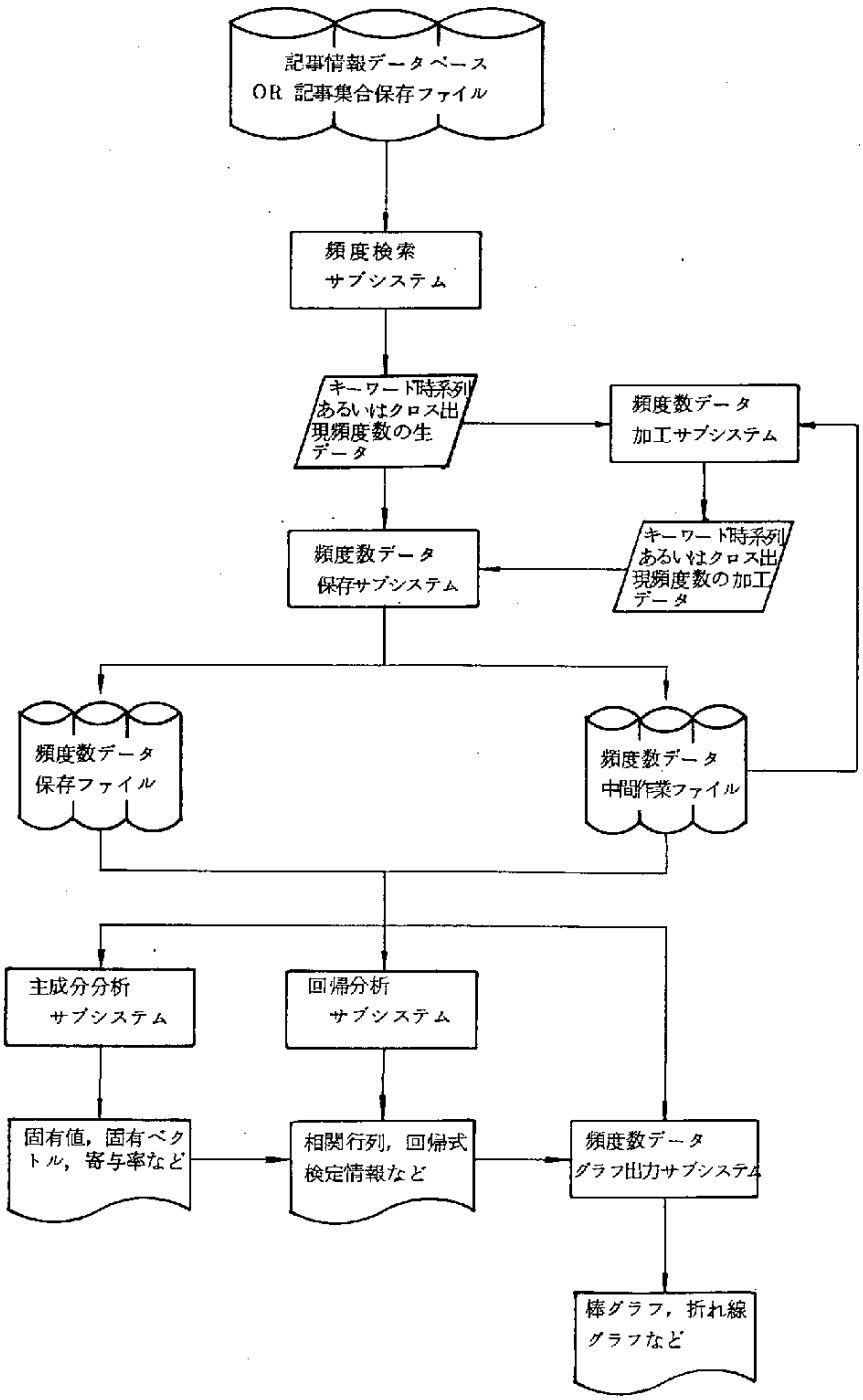


図 2.3－2　サブシステム間の情報関連図

2.3.4　システムの概要

各サブシステム単位に，サブシステムの機能，入力情報，出力情報，作成ファイル等の概要について述べる。

定量化分析システムの入出力情報の一覧は，図2.3－3に示す通りである。

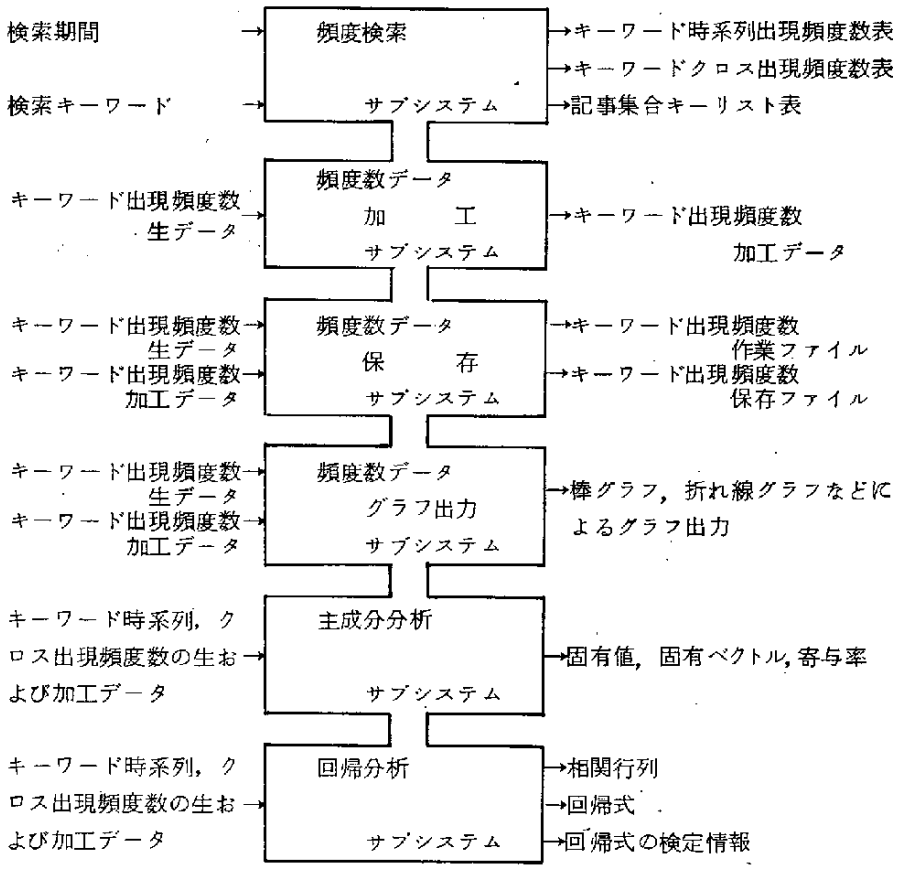


図2.3－3　定量化分析システムの入出力

(1) 頻度検索サブシステムの拡充

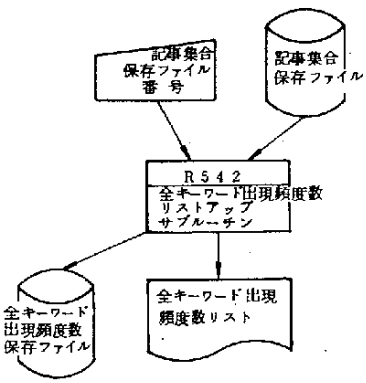
頻度検索サブシステムについては，データベース検索システム基本設計書の中で，キーワード出現頻度数の時系列検索機能，クロス検索機能を設定した。ここでは，さらに多方面からキーワード出現頻度数の生データを集めるため，頻度検索の機能を次のように拡充する。

i) 検索対象ファイルを記事インデックスのソース・データベースだけでなく，検索期間を絞り込んだ０１記事保存ファイルおよび１次検索，論理検索，２次検索で作成した記事保存ファイル（０２～）も検索対象とする。

ii) 記事保存ファイルを対象として，その中の記事インデックスが持つキーワードをリストアップし，出現頻度数を出し，保存ファイル間の出現頻度パターンの比較もできるようにする。

(a) 拡充機能

① 全キーワード出現頻度数リスト・アップ機能

処理フロー	ブロック名	機能
		記事集合保存ファイルを対象として，その中に存在する全キーワードと出現頻度数をリストアップする。
	R542	ディスプレイから記事集合保存ファイル番号を受信し，出現頻度数の多いキーワードから順次全キーワードについて出現頻度数を漢字プリンタにリストアップする。比較のためデータをファイルに蓄積する。

(b) 頻度検索サブシステムの入力情報

拡充機能部分の入力情報は，記事集合保存ファイル番号，検索結果保存指示データ，検索終了データなどである。

(c) 頻度検索サブシステムの出力情報

拡充機能部分の出力情報は，1つの保存集合に対する全キーワード出現頻度数リストである。

全キーワード出現頻度数リストは，出現頻度数の大きい順に存在するキーワードと出現頻度数を見易い一定の書式で出力する。

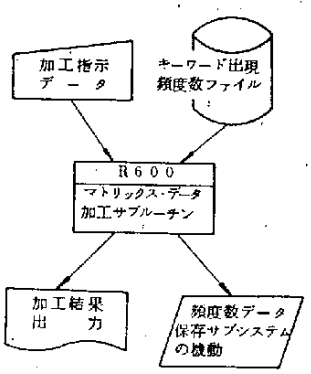
(d) 頻度検索サブシステムで作成するファイル

全キーワードの出現頻度数を出現頻度の高い順にキーワード名とともに蓄積する中間作業ファイルが，当サブシステムでは作成される。キーワード情報の頭には，記事集合保存ファイル番号及び保存ファイル名も併せて登録される。

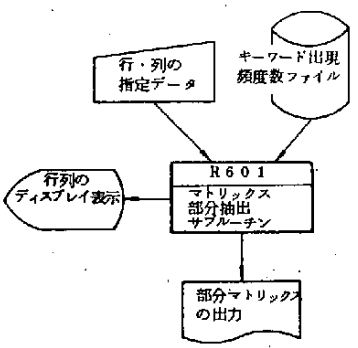
(2) 頻度数データ加工サブシステム

(a) 頻度数データ加工サブシステムの機能

① マトリックスデータ加工機能

処理フロー	ブロック名	機能
		キーワードの時系列あるいはクロス出現頻度データ(マトリックス)にいくつかの加工を加えて加工データを作るとともに結果を出力する。
	R600	加工指示データをディスプレイから受信し，マトリックス型のデータとして保存されているキーワード出現頻度数を行単位あるいは列単位で増減率，構成比，四則演算を行ったりして加工する。 加工したデータを保存したい場合は，頻度数データ保存指示データを受けて，保存システムを機動する。加工終了データがくるまでは，データ加工を続ける。

② マトリックス部分抽出機能

処理フロー	ブロック名	機能
		頻度検索で求めた時系列，あるいはクロス出現頻度数の生データあるいは加工データから分析用の部分マトリックスを作成する。
	R601	主成分分析，クラスター分析，回帰分析用のデータを頻度検索で得た出現頻度数のデータから作成。オリジナル行列の行と列のインフォーメーションをディスプレイに表示し，必要な行と列を指定して部分マトリックスを構成する。

(b) 頻度数データ加工サブシステムの入力情報

当サブシステムの入力情報は，加工指示データ，行，列の指定データ頻度数ファイル番号などである。

当サブシステムでは，キーワード出現頻度数ファイルからオリジナルデータを読み込んで，データ加工，マトリックス編集を行う。

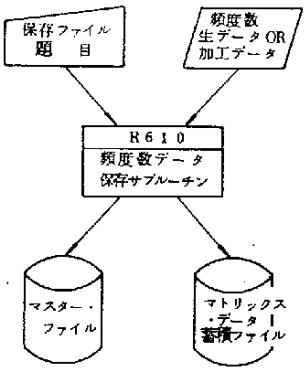
(c) 頻度数データ加工サブシステムの出力情報

当サブシステムの出力情報は，加工データ，編集マトリックスを元の頻度数データと同じようなマトリックスの形で漢字プリンタに表示する。

データ加工，マトリックス編集を容易にするため，行および列のインフォメーションをできる限りディスプレイ表示する。

(3) 頻度数データ保存サブシステム

(a) 頻度数データ保存サブシステムの機能

処　理　フ　ロ　ー	ブロック名	機　　能
		頻度検索で得られたキーワード出現頻度数の生データおよび加工サブシステムで得られた加工データをマトリックス形式で蓄積する。
	R610	頻度数データ保存指示データを受信したら，結果保存ファイル番号を示し，保存ファイルの題目を入力させる。題目入力後保存ファイル番号，題目，マトリックスの行および列の大きさ，行および列の名前，マトリックス・データ蓄積ファイル内のデータ開始アドレスと終了アドレスをマスター・ファイルに保存する。マトリックス・データは一次元データに直して適切な個数ずつファイルに書き込む。

(b)　頻度数データ保存サブシステムの入力情報

当サブシステムの入力情報は，保存ファイルの内容を示すためのメモである（２０字以内）。

(c)　頻度数データ保存サブシステムで作成されるファイル

当サブシステムでは，頻度数データ保存のためマスターファイルとマトリックス・データ蓄積ファイルが作成される。

マスター・ファイルの内容は，保存ファイル番号，保存ファイル題目，マトリックスの行数，列数，マトリックス・データ蓄積ファイルの開始アドレス，終了アドレス，行および列のキーワードを保存する。フォーマットは次の通りである。

ファイル番号 ３バイト	保存ファイル題目 ４０バイト	行　数 ３バイト	列　数 ３バイト	開始アドレス １０バイト	終了アドレス １０バイト

行のキーワード1 30バイト	行のキーワード2 30バイト	.....

列のキーワード1 30バイト	列のキーワード2 30バイト	.....

キーワードのためには，行Max 100個，列Max 100個の6,000バイト用意する。

もし，行および列のキーワードの個数に上限値を設定したくない場合は，マトリックス・データの蓄積と同様の方法で，行・列名ファイルを別途作成し，マスターファイルにはその開始アドレスと終了アドレスを持つ。

マトリックス・データ蓄積ファイルは，出現頻度数のデータを行数×列数の1次元数値群に直して，500個程度ずつファイル蓄積する。例えば，30行×40列のマトリックスの場合，ファイルは下記のようになる。

開始アドレス

1番目			500番目
D(1, 1) 10バイト	D(2, 1) 10バイト	.....	D(20, 17)

501番目			1000番目
D(21, 17) 10バイト	D(22, 17) 10バイト	.....	D(10, 34) 10バイト

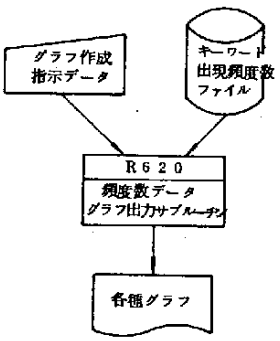
終了アドレス

1001番目				1500番目
D(11, 34) 10バイト	D(12, 34) 10バイト	.....	D(30, 40)	(空　　白)

1レコードに書き込むデータ個数はファイル処理効率の良好な値を選定すればよい。

(4) 頻度数データグラフ出力サブシステム

(a) 頻度数データグラフ出力サブシステムの機能

処理フロー	ブロック名	機能
		キーワード出現頻度数の生データ及び加工データ，各分析サブシステムの分析結果をビジュアルなグラフで漢字プリンタあるいはX－Yプロッタに出力する。
	R620	グラフ出力の指示データをディスプレイから受け取ったら，グラフの種類を選択させ，各グラフ作成法の指示に従ってデータを指定して，保存ファイルなどからデータを読み込み，漢字プリンタあるいはX－Yプロッタに出力する。

(b) 頻度数データグラフ出力サブシステムの入力情報

当サブシステムの入力情報は，グラフ選択データ，グラフ作成指示データ，グラフ出力終了データである。

グラフ作成指示データに従って，キーワード出現頻度数ファイルあるいは各分析システムの分析結果から必要なデータを読み込む。

(c) 頻度数データグラフ出力サブシステムの出力情報

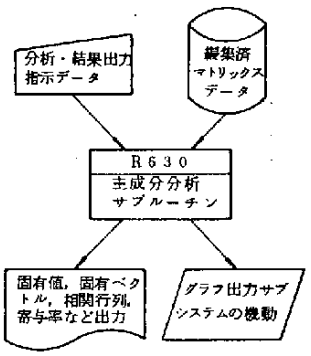
当サブシステムでは，ユーザの指示に従って①棒グラフ，②折れ線グラフを出力する。

棒グラフは各種キーワードの出現頻度数の特性比較などに用いられる。折れ線グラフはキーワード出現頻度数の時系列変化に用いる。

(5) 主成分分析サブシステム

(a) 主成分分析サブシステムの機能

処理フロー	ブロック名	機能
		時系列あるいはクロスのキーワード出現頻度数の生データあるいは加工データに基づいて主成分分析を行う。



R630	頻度数データ加工サブシステムのマトリックス部分抽出機能あるいは結合機能を用いて，必要であれば分析に用いるためのマトリックスを編集しておく。 準備したマトリックスから出発して主成分分析を行い，相関マトリックス，固有値，固有ベクトル，寄与率を求め，漢字プリンタに出力する。指示データに従ってグラフ出力も行う。

(b) 主成分分析サブシステムの入力情報

当サブシステムでは，加工サブシステムの部分抽出機能あるいは結合機能を用いて，あらかじめ編集したマトリックスからデータを読み込み主成分分析を行う。主成分分析は，時系列頻度数データあるいはクロス頻度数データいずれをも対象として行える。

主成分分析を行うためには，系列数およびサンプル数が入力データとして必要であるが，マトリックス編集の段階で行数，列数が与えられるので，それを用いる。

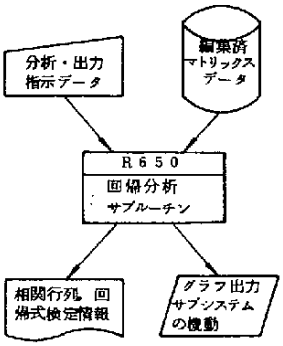
結果出力に関しては，いくつか指示データを入力してプリンタ出力，グラフ出力を行うようにする。

(c) 主成分分析サブシステムの出力情報

主成分分析の結果として，相関行列，固有値，固有ベクトル，寄与率などをプリンタ出力する。

(6) 回帰分析サブシステム

(a) 回帰分析サブシステムの機能

処　理　フ　ロ　ー	ブロック名	機　　能
		時系列あるいはクロスのキーワード出現頻度数の生データあるいは加工データに基づいて回帰分析を行う。
	R650	加工サブシステムを用いてマトリックスを編集しておく。準備したマトリックスをオリジナル・データとして，時系列データの場合は，回帰期間を与え系列名を指定しながら回帰分析を行う。 ディスプレイに出力した回帰式と検定情報を見て，漢字プリンタへの出力，グラフ出力などの指定を行う。変数，期間，データファイルなどを変更して終了データがくるまで動作を続ける。

(b)　回帰分析サブシステムの入力情報

当サブシステムでの入力情報は，キーワード頻度数保存ファイル指定データ，回帰期間，説明変数，被説明変数の系列名データ，分析終了データなどである。

当サブシステムで利用する入力ファイルは，他の分析サブシステムと同様，加工サブシステムのマトリックス抽出機能で編集しておく。

(c)　回帰分析サブシステムの出力情報

回帰分析サブシステムでは，回帰結果を相関係数，ダービン・ワトソン比，標準偏差，t−値などの検定情報とともにディスプレイ上に示す。

ディスプレイに出力された結果を判断して，漢字プリンタへ実績値・計算値の比較，相関行列，回帰式，検定情報を出力し，推定結果のグラフを出力する。

2.4 データ整備

海外で発行されている雑誌等の主要５０紙誌を情報源として，第２次石油危機以後の１９７９年１月から５カ年分を対象に，エネルギー関連の記事情報を抽出してインデックスを作成し，更にキーワード付けを行った。

2.4.1 データ整備の方法

抽出した記事について記事全体を把握し，各記事に対して内容を４０字以内にまとめた記事インデックスを作成し，掲載年月日，掲載紙面あるいはページ，紙誌名略号，記事行数を付加してデータを作成しコーディング作業を行った。さらに記事内容にふさわしいキーワードを海外会社，海外団体，項目，品目，地域等の大分類に従って選択し，検索あるいは定量化解析に利用するためキーワード付けを行った。これらのデータは，コーディング後パンチされ，磁気テープ化された。

2.4.2 整備したデータの量

今回のデータ整備では，５７年度整備したデータに加えて，７９年１月から８３年１２月まで５カ年間のエネルギー関連記事のインデックスを作成した。

データベースに登録したデータ件数は，総計３４,５１２件であり，年別の内訳は次のとおりである。

年	件　数
７９年	9,９０３
８０年	8,６５４
８１年	7,９６３
８２～８３年	7,９９２
計	３４,５１２

作成した記事インデックスの例は，データベースの検索出力例に示すとおりである。

出力例１　記事インデックス検索結果

記事インデックス	年月日	紙誌名	紙面・ページ	行数	記事番号
米国下院予算委員会、海洋石油・天然ガス開発優先――石炭関発予算削減	830622	WJ	006	139	8306224005
エネルギー戦略の崩壊で米国の合併や公開買付攻撃ちつく	830611	ECO	079		8306114005
パーカーＮＣＡ会長、米石炭業界の将来に明るい見通し――実質年４％成長	821025	CN	001		8210254027
米政府、オクシデンタル社のシベリア石炭パイプライン建設を阻止する意向	821008	WJ	005	80	8210084010
米国合成燃料公社、石炭含む１１プロジェクトに対する補助金検討を約束	820927	CN	001		8209274016
米西部９州の知事が産炭地貸与計画に関しワット内務長官に抗議文書送る	820906	IHT	003	98	8209064020
ポーランド、石炭輸出で米国など西側輸出国との競争に直面	820831	FT	002	88	8208314007
米国上院エネルギー・資源委、石炭パイプライン法案を可決	820811	CWI	008		8208114009
米上院エネルギー委の石炭パイプライン法採択で鉄道の石炭輸送独占に幕	820810	WJ	028	100	8208104009
米国上院エネルギー委、圧倒的多数でスラリー炭パイプライン法案を可決	820809	CN	001		8208094022
米下院公共事業委、石炭スラリー・パイプライン建設促進法案を可決	820730	WJ	038	55	8207304002
米石炭輸出の増強には港湾設備の近代化が不可欠――ヘリテージ財団が調査	820712	CN	003		8207124032
米国産石炭価格、現政策のままでは２１世紀までに上昇し州・民間業者に利益大	820621	CW	002		8206214037
米国の３上院議員、輸出向け石炭の鉄道運賃規制撤廃反対で業界支持を表明	820621	CN	001		8206214032
米国石炭業界専門家、州際商業委の鉄道運賃決定は一貫性に欠けると非難	820614	CW	008		8206144031
米国石炭業界、エネルギー生産・消費新税に断固反対――ＮＣＡ会長が議会証言	820614	CN	001		8206144028
米石炭業界代表、議会に７６年石炭資産借用法の柔軟な改正を強く要請	820614	CN	001		8206144027
米石炭協会、将来のエネルギー対策として石炭利用拡大の重要性を強調	820531	CN	002		8205314004
米国石炭輸送合理化連合、議会で石炭パイプライン輸送確立の必要性を強調	820517	CN	001		8205174008
米国石炭協会、８３年度の石炭関連予算削減は石炭消費拡大に打撃と批判	820503	CN	002		8205034035
米政府、ワイオミングの産炭地貸与入札の対象から６カ所を削除	820419	CN	001		8204194029
米の石炭積み出し増整備計画、輸出ブームの衰退で規模縮小	820330	WJ	027	201	8203304003
米石炭業界、石炭輸送の鉄道独占打破でスラリー輸送への法的整備を要求へ	820305	CL	008		8203054026
米の炭鉱事故死者、年頭以来３１名――政府、保安要員削減計画を再考	820227	ECO	041		8202274008
アイオワ商業委、アイオワ・パブリック・サービスの自社炭で適正価格指導	820222	CW	004		8202224034
米政府と石炭業界、大気汚染防止法緩和法案を共同支持――石炭消費拡大への道	820208	CW	007		8202084036
米Ｊ＆Ｌコール・アソシエーツ、州際商業委に石炭鉄道輸送契約の公表を要求	820208	CW	002		8202084035
米合成燃料公社の石炭関連事業選択、市場・政治的要因がカギ――８計画が存続	820125	CW	005		8201254025
仏、８１年の石炭輸入を２００８万ｔに減少――ＥＣ域外からは米国炭が最大	820118	EP	013		8201184028
米内務省、石炭業界の意向にそって産炭地貸与条件を大幅改正	811221	CW	003		8112214018
米エネルギー省、石油・天然ガス火力発電所の石炭転換適用除外を拡大へ	811214	CW	004		8112144033
ＥＰＡ、新法制導入の石炭火力の大気汚染防止規制適用除外延長で議会承認期待	811214	CW	002		8112144032
米上院エネルギー委、世界石炭市場における米の役割について聴聞会２日間延長	811125	CW	007		8111254031
石炭パイプラインの決定は各州で――米内務省、連邦政府の積極的支持に否定的	811123	CW	006		8111234024

米議会、石炭輸出・船腹問題で公聴会――対欧輸出増加に否定的見解も	811118	CWI	007		8111184028
米石炭輸送業界、鉄道輸送契約規制で様々な論議――ICC鉄道は慎重対応	811116	CW	002		8111164033
米司法省、石炭輸出増強施設増強計画への法的規制の緩和を予測	811111	CWI	008		8111114034
米州際商業委、石炭輸送業者の鉄道運賃対策・ガイドラインで決定へ	811102	CW	001		8111024026
米の石炭消費量、82年までに7600万t増加へ――エネルギー省予測	811026	CW	007		8110264028
米産業用石炭消費、ガス規制緩和政策推進で大幅増――業界が予測	811026	CW	003		8110264027
米石炭州分離税12．5%規制は逆効果――資源研究所員が批判	811026	CW	001		8110264026
米アパラチア炭の需要は天然ガス価格統制撤廃がカギ――90年までに倍増も	811019	CW	004		8110194021
米連邦エネルギー規制委決定、石炭市場・炭質などの定義が不明確との批判も	811019	EW	009		8110194019
インドネシア、外国企業と共同で石炭開発――狙いは石油輸出力の延命	811012	AWJ	003	150	8110124002
米会計検査院、緊急時における石炭への転換計画の実効性について批判	811005	CW	002		8110054041
レーガン米政権のエネ省燃料転換局位廃案、石炭転換計画企業に大きな痛手	811005	EW	001		8110054035
米内務省、カーター前政権が禁止した国立公園近くの石炭開発を再検討	810922	WJ	009	69	8109224001
米EPA、石炭燃焼プラントなどの煤ガス規制基準変更の審査時間を半減へ	810914	CW	002		8109144044
米商務省、石炭スラリー推進問題で対立――決定はホワイトハウスに一任	810914	CW	001		8109144043
米電力業界、石炭使用増加促進の燃料使用法についてエネ省に陳情	810824	EW	007		8108244041

注記

石炭×エネルギー政策F×米国　全177件

出力例２　記事インデックス検索結果（全キーワード付き）

出力形式：記事インデックス　年月日　紙誌名　紙面・ページ　行数　記事番号　　全キーワード

フランスの石炭政策、経済性と地域政策の狭ばさみ　830531　LM　044　200　8305314008
（海外団体）　フランス石炭公社　フランス政府　（品目）　炭鉱　石炭　（業界）　石炭業界Ｆ　（項目）　助成金Ｆ　産業政策Ｆ　エネルギー政策Ｆ　地域開発Ｆ
（項目）　需給見通しＦ　販売見通しＦ　原価計算Ｆ　（海外地域）　フランス　（紙誌名略号）　ＬＭ

フランス労働組合型など、フランス原発計画の下方修正に反発　830517　LM　045　130　8305174013
（海外団体）　フランス労働組合型　フランス政府　（品目）　原子力発電　天然ガス　（項目）　需要見通しＦ　経済成長Ｆ　原子力政策Ｆ　エネルギー政策Ｆ
（項目）　原子力開発Ｆ　電源開発Ｆ　省エネルギーＦ　（海外地域）　フランス　（紙誌名略号）　ＬＭ

フランス、新税導入で石油製品価格が値上がり　830512　LM　028　10　8305124006
（海外団体）　フランス政府　（品目）　石油製品　ガソリン　灯油　軽油　（項目）　エネルギー政策Ｆ　税制改正Ｆ　仏税政策Ｆ　価格政策Ｆ　値上げＦ
（海外地域）　フランス　（紙誌名略号）　ＬＭ

フランス、４月に創設された補助税の代わりに内国税体系を改訂　830512　LM　022　38　8305124005
（海外団体）　フランス政府　（品目）　石油製品　（項目）　価格政策Ｆ　価格動向Ｆ　税制改正Ｆ　エネルギー政策Ｆ　（海外地域）　フランス　（紙誌名略号）　ＬＭ

フランス、産業用および暖房用に石炭消費を奨励　830511　CWI　007　8305114009
（海外団体）　フランス電力公団　フランス政府　（品目）　石炭　（項目）　エネルギー政策Ｆ　産業界Ｆ　補助金Ｆ　需要動向Ｆ　燃料転換Ｆ　（海外地域）　フランス
（紙誌名略号）　ＣＷＩ

フランス、４月はガソリン価格の値下げなし　830328　LM　017　40　8303284002
（海外団体）　フランス政府　ＯＰＥＣ　（品目）　原油　石油製品　ガソリン　灯油　軽油　（項目）　産油国　税制改正Ｆ　消費税Ｆ　価格政策Ｆ　エネルギー政策Ｆ
（項目）　価格動向Ｆ　仏税政策Ｆ　値下げＦ　（海外地域）　フランス　（紙誌名略号）　ＬＭ

フランス、ＯＰＥＣ原油値下げにもかかわらず４月石油製品値下げ見送りか　830317　LM　034　20　8303174002
（海外団体）　ＯＰＥＣ　フランス政府　（品目）　ガソリン　原油　石油製品　（項目）　価格政策Ｆ　エネルギー政策Ｆ　値下げＦ　（海外地域）　フランス
（紙誌名略号）　ＬＭ

フランス、理論的にはガソリン価格１１当たり１５サンチームの値下がり　830316　LM　043　80　8303164013
（海外団体）　ＯＰＥＣ　フランス政府　（品目）　ガソリン　灯油　軽油　原油　石油製品　（項目）　価格政策Ｆ　エネルギー政策Ｆ　産油国　価格動向Ｆ　値下げＦ
（海外地域）　フランス　（紙誌名略号）　ＬＭ

フランス、３月１０日に再び石油製品価格引き下げ　830303　LM　032　43　8303034002
（海外団体）　フランス政府　（品目）　軽油　石油製品　ガソリン　灯油　（項目）　価格政策Ｆ　値下げＦ　エネルギー政策Ｆ　（海外地域）　フランス
（紙誌名略号）　ＬＭ

フランスのキャルモー炭鉱スト、国内炭生産水準をめぐる論議が再燃　830303　LM　001　180　8303034001
（海外団体）　フランス石炭公社　（品目）　石炭　（項目）　需要見通しＦ　雇用Ｆ　ストＦ　生産計画Ｆ　エネルギー政策Ｆ　労働問題Ｆ　（海外地域）　フランス
（紙誌名略号）　ＬＭ

フランス、スポットもの価格などを織り込んで石油製品最高価格を引き下げ　830205　LM　036　51　8302054001
（海外団体）　フランス政府　（品目）　灯油　石油製品　軽油　ガソリン　（項目）　エネルギー政策Ｆ　価格政策Ｆ　値下げＦ　（海外地域）　フランス
（トピック）　スポット価格Ｆ　（紙誌名略号）　ＬＭ

フランス、石油製品１１当たり３サンチーム値下げ　830203　LM　032　51　8302034012
（海外団体）　フランス政府　（品目）　石油製品　灯油　ガソリン　軽油　（項目）　値下げＦ　価格政策Ｆ　エネルギー政策Ｆ　（海外地域）　フランス
（紙誌名略号）　ＬＭ

注記
（西独＊産業政策）＋（＊フランス＊エネルギー政策Ｆ）　全３５３件

出力例３　キーワード月別頻度検索結果

検索キーワード：　エネルギー外交F

7901	7902	7903	7904	7905	7906	7907	7908	7909	7910
24	34	48	31	22	34	21	27	30	47
7911	7912	8001	8002	8003	8004	8005	8006	8007	8008
30	31	31	23	47	36	27	27	24	17
8009	8010	8011	8012						
11	25	21	23						

検索キーワード：　原油

7901	7902	7903	7904	7905	7906	7907	7908	7909	7910
197	225	286	181	174	202	154	104	99	176
7911	7912	8001	8002	8003	8004	8005	8006	8007	8008
131	211	159	147	138	126	146	124	104	99
8009	8010	8011	8012						
103	78	78	101						

出力例４　キーワード国別頻度検索結果

検索キーワード：　産業政策F

米国	英国	日本	西独	フランス	スウェーデン	イタリア	ベルギー	オランダ	スペイン
41	67	1	13	27	0	2	2	2	0

検索キーワード：　石油

米国	英国	日本	西独	フランス	スウェーデン	イタリア	ベルギー	オランダ	スペイン
1995	647	83	263	183	0	60	11	50	14

3. 文章情報解析手法の研究

3. 文章情報解析手法の研究

3.1 世界エネルギー情報によるキーワード出現頻度数分析

3.1.1 分析実験の概要

データベース化された大量の文章情報を分析し，そこに潜在する徴候的情報，基調的変化を把握することはきわめて重要と考えられる。おかしいと気付けば，我々は持てるありとあらゆる手段を駆使して分析にあたることができるわけであるが，問題はおかしいと気付くきっかけを与える徴候的情報をいかに得るかということである。データベース化された大量の文章情報の処理によって得られる情報は，このような動きをとらえる手段を提供する可能性を持っていると言えよう。

本研究では，１９７９年1月から現時点まで5年間で約５０種の海外紙誌に掲載されたエネルギー関連の記事を対象として，４０字以内の記事インデックスの作成とキーワード付けを行い記事情報のデータベース化を行った。１９７９年から5年間は，第２次石油危機を契機としてエネルギー情勢がめまぐるしく変転した時代にあたっており，作成した文章情報データベースはその意味では大変興味深い実験材料となっている。そこで，本節では記事インデックスに付されたキーワードの出現頻度を様々な角度から分析し，文章情報総合解析システムが装備すると有効と考えられる定量化分析機能の検討を行った。

分析実験の概要を表3.1－1に示す。実験では，表3.1－1に示した各種のキーワード出現頻度数データに基づいてＯＰＥＣ加盟国を中心とした分析を行ったが，紙数の都合で本論ではＯＰＥＣの盟主サウジアラビアと石油供給過剰の影響を最も強く被ったナイジェリアの分析結果を中心として報告する。表3.1－2には，分析の対象となったキーワードをその略号（括弧内）とともに示している。以後の図中では，表3.1－

表3.1－1　分析実験の概要

＜オリジナル・データ＞	＜分析手法＞	＜出力結果＞
年別 ＯＰＥＣ加盟国別 各種キーワード出現頻度数	（クロスデータ） 頻度数分析	ＯＰＥＣ加盟国の分布パターンと時系列変化
	時系列プロット	頻度数の増減とキーワード間のタイムラグ
	主成分分析	ＯＰＥＣ加盟国の結びつき
	クラスター分析	ＯＰＥＣ加盟国の類似度
月別 ＯＰＥＣ加盟国別 原油値上げ・値下げ キーワード出現頻度数	（時系列データ） 時系列プロット	原油価格の変化とキーワードの出現頻度数 波及効果
四半期別 各種キーワード出現頻度数 （データベース別）	（時系列データ） 時系列プロット	データベース比較
	主成分分析	キーワードの結びつき
	回帰分析	キーワード間の時系列相関
年別 先進4カ国別 エネルギー源別 記事出現頻度数	（クロスデータ） レーダー・チャート	エネルギー源への関心と時系列変化

2に示した略号がすべてのキーワードを表示するため使用されている。

表3.1－2　検索に使用したキーワード

	エネルギー資源	エネルギー関連	生産関連	価格関連	経済関連	軍事・政治	採鉱・設備	産油国関連
1	石油（石油）	エネルギー政策（エ政）	生産計画（生計）	価格政策（価政）	経済動向・経済見通し・経済計画（経動）	兵器・国際紛争・戦争（戦争）	石油探鉱・油田（油田）	ＯＰＥＣ（ＯＰ）
2	原油（原油）	エネルギー外交（エ外）	生産動向（生動）	価格動向（価動）	経済外交・経済協力（経外）	反政府運動・クーデター・ゲリラ・過激派暴動（反政）	製油所・パイプライン（製油）	石油会議（会議）
3	石油製品（製品）	エネルギー開発（エ開）	生産見通し（生見）	価格見通し（価見）	インフレ（イン）	政権交代・政治・政治体制（政治）		産油国（産国）
4	天然ガス・LNG・LPG（ガス）	エネルギー問題（エ問）	増産（増産）	値上げ（値上）	オイルマネー（マネ）			石油消費国（消国）
5	電力・原子力（電力）		減産（減産）	値下げ（値下）	対外債務・借款（債務）			石油業界・メジャー（業界）
6	石炭（石炭）			価格交渉（価交）	産業政策（産政）			
7				プレミアム販売（プレ）				
8				商品市況・市場動向（市況）				

（）内はキーワードの略号

3.1.2　キーワード出現頻度数の時系列分析

(1)　キーワード出現頻度数にみる話題の推移

この分析実験では，表3.1－2で示したキーワードの出現頻度数をＯＰＥＣ加盟各国について年単位で得たものをまず用いている。図3.1－1には，サウジアラビアとナイジェリアについて，各キーワードの出現頻度数をそれぞれの国名キーワードの出現頻度数で除して百分比としたものをプロットしている。各年次のキーワードを百分比の大小順に並べて点線でプロットし，翌年のキーワードの百分比をそのキーワード順に実線でプロットすることにより話題の推移を見られるようにしたものである。点線よりも実線がへこんでいるキーワードは話題にのぼる度合が小さくなったことを示しており，逆にふくらんでいるキーワードは話題にのぼる度合が大きくなったことを示している。

サウジアラビアの場合，７９年から８０年へかけてあまり大きな変化はなく増産とスウィング・プロデューサとしての生産計画に関する話題がいく分減っている。８０年から８１年にかけては，値上げが減る一方で減産，値下げ，石油会議のキーワードが増大しており，話題の変動がかなり激しくなっている。８１年，８２年，８３年へとかけては減産のキーワードが減り，生産調整を示す生産計画のキーワードが増えながら，石油業界・メジャーの圧力も受けて，８３年３月の原油価格大幅値下げへ推移していく様子がよくわかる。

ナイジェリアの場合は，サウジアラビアよりも早く，７９年から８０年への変化の中で値上げのキーワードは減少し，減産あるいは値下げのキーワードが増加し，話題の大きな変化が示されている。８１年にかけては，ナイジェリアが他のＯＰＥＣ諸国に先駆けて大幅値下げを行ったため，価格政策，値下げのキーワードが急増している。８２年は，石油供給過剰のあおりを被ったナイジェリアの生産動向が注視され，石油業界・メジャーも強い圧力をかけたので生産動向と石油業界

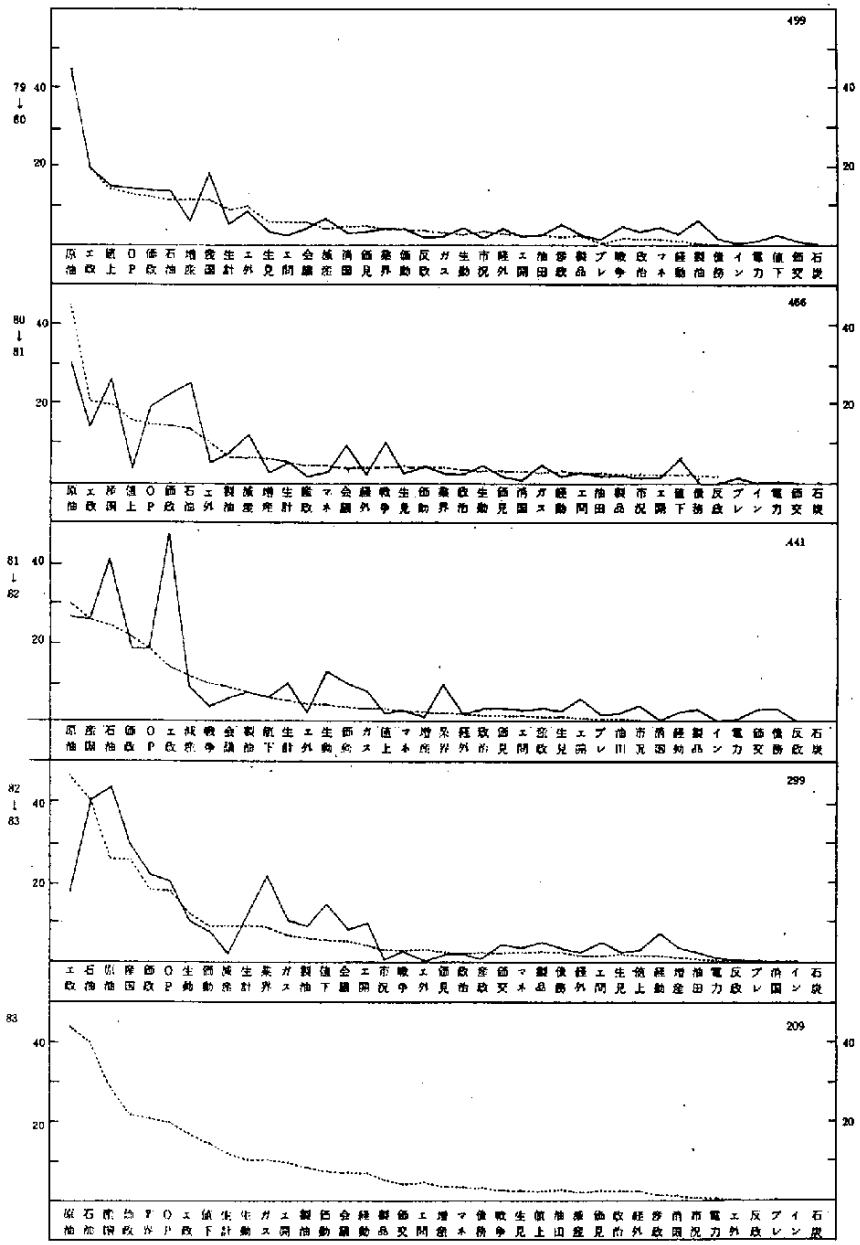


図 3.1－1　キーワード出現頻度数にみる話題の推移 (1)

― サウジアラビア ―

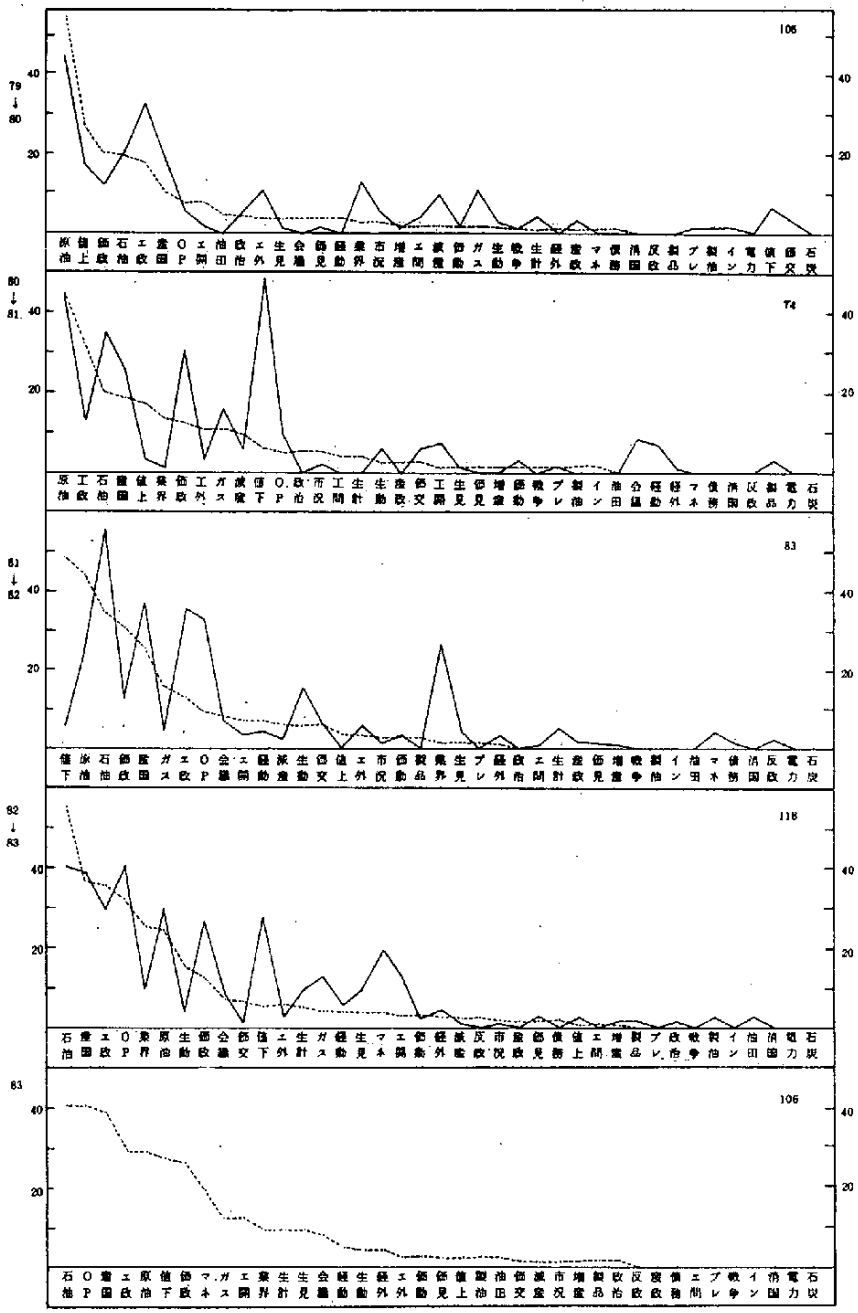


図 3.1－1　キーワード出現頻度数にみる話題の推移(2)

－ナイジェリア－

（注１）点線は当該年，実線は翌年
（注２）右上の数字は国名キーワードの出現頻度数

のキーワードが増大した。８３年はナイジェリアのオイルマネーと財政ひっ迫が話題となり，８３年３月の原油価格大幅値下げに際してもナイジェリアの行動は注視されたため，値下げとオイルマネーのキーワードが伸びている。

この他にも，まだまだ細かい話題の推移をとらえることはできるが，このようにキーワードの出現頻度数を大小順に並べて話題の推移を時系列に分析していく手法は，様々な事象の変化を大づかみにとらえる上で有効と考えられる。

(2) キーワード出現頻度の順位変化

図3.1－1では，キーワードの出現頻度数を年別に大小順に並べて話題の推移を追ってみたが，図3.1－2では，図3.1－1で得られたキーワードの出現頻度の順位変化を上位１０個のキーワードに絞って追いかけている。□で囲まれたキーワードは，イラン革命に端を発する第２次石油危機の勃発で原油価格が急騰した７９年において上位１０の出現頻度を示したものである。[ ]で囲まれたキーワードは，石油供給過剰で原油価格の大幅値下げをせざるを得なかった８３年に，７９年に上位１０を占めたキーワード以外で新たに上位１０に入ったキーワードを示している。(　)で囲まれたキーワードは，８０年から８２年の途中年次においてのみ上位１０に加わったキーワードを示している。

原油，石油，産油国，ＯＰＥＣ，エネルギー政策，価格政策の６個のキーワードは上位１０の中で上下の変動はあるが，サウジアラビア，ナイジェリアともに5年間変化なく表われているので，ＯＰＥＣ加盟国に共通でしかも時系列の変動も少ないキーワードと言うことができよう。

値上げ，減産，生産動向，石油会議，石油業界・メジャー，値下げのキーワードは，時系列の変動は激しいが，サウジアラビア，ナイジェリアともほぼ同様の変化を示しており，ＯＰＥＣ加盟国がある程度

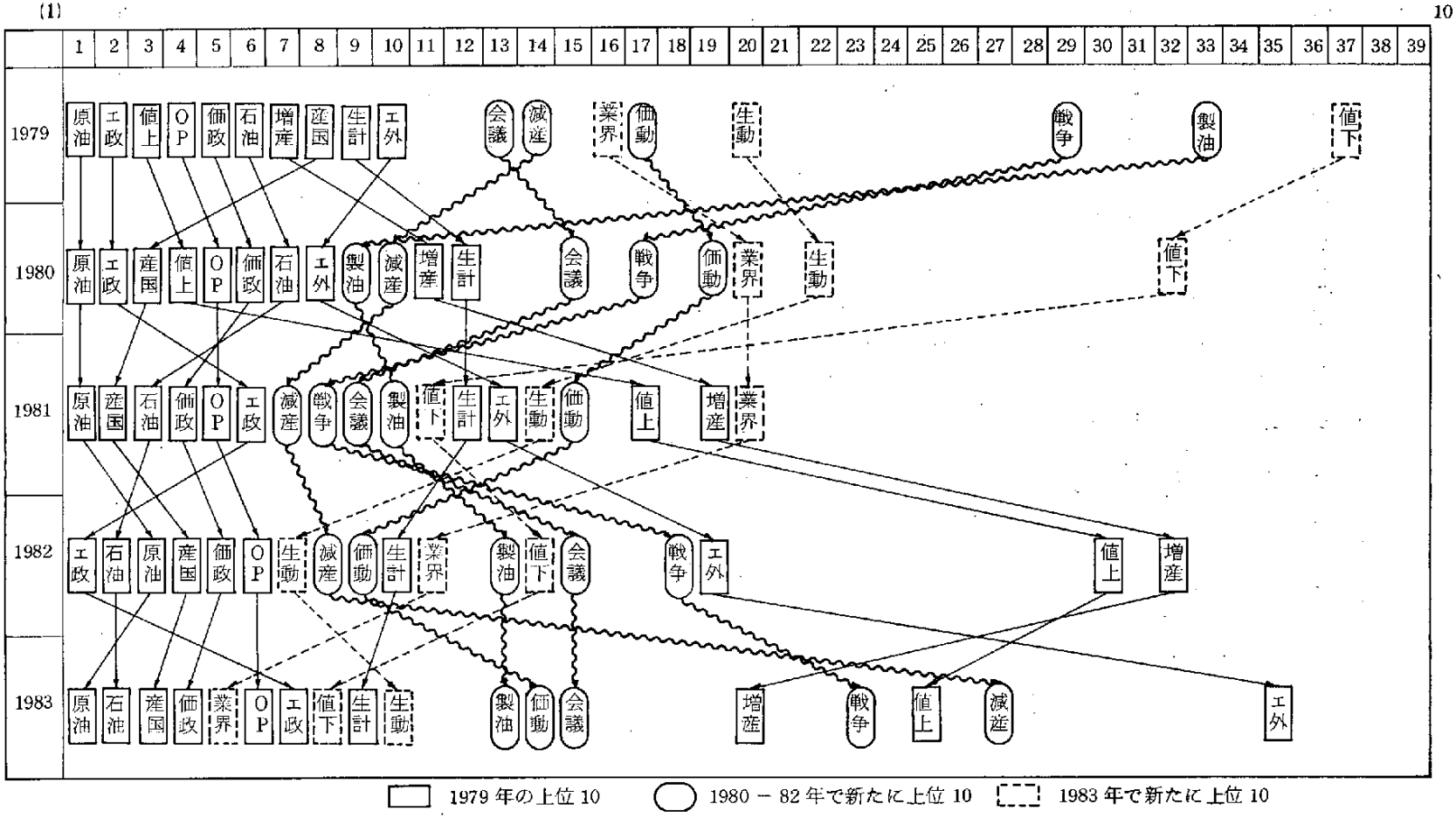


図 3.1－2　キーワード出現頻度順位変化(1)― サウジアラビア ―

(2)

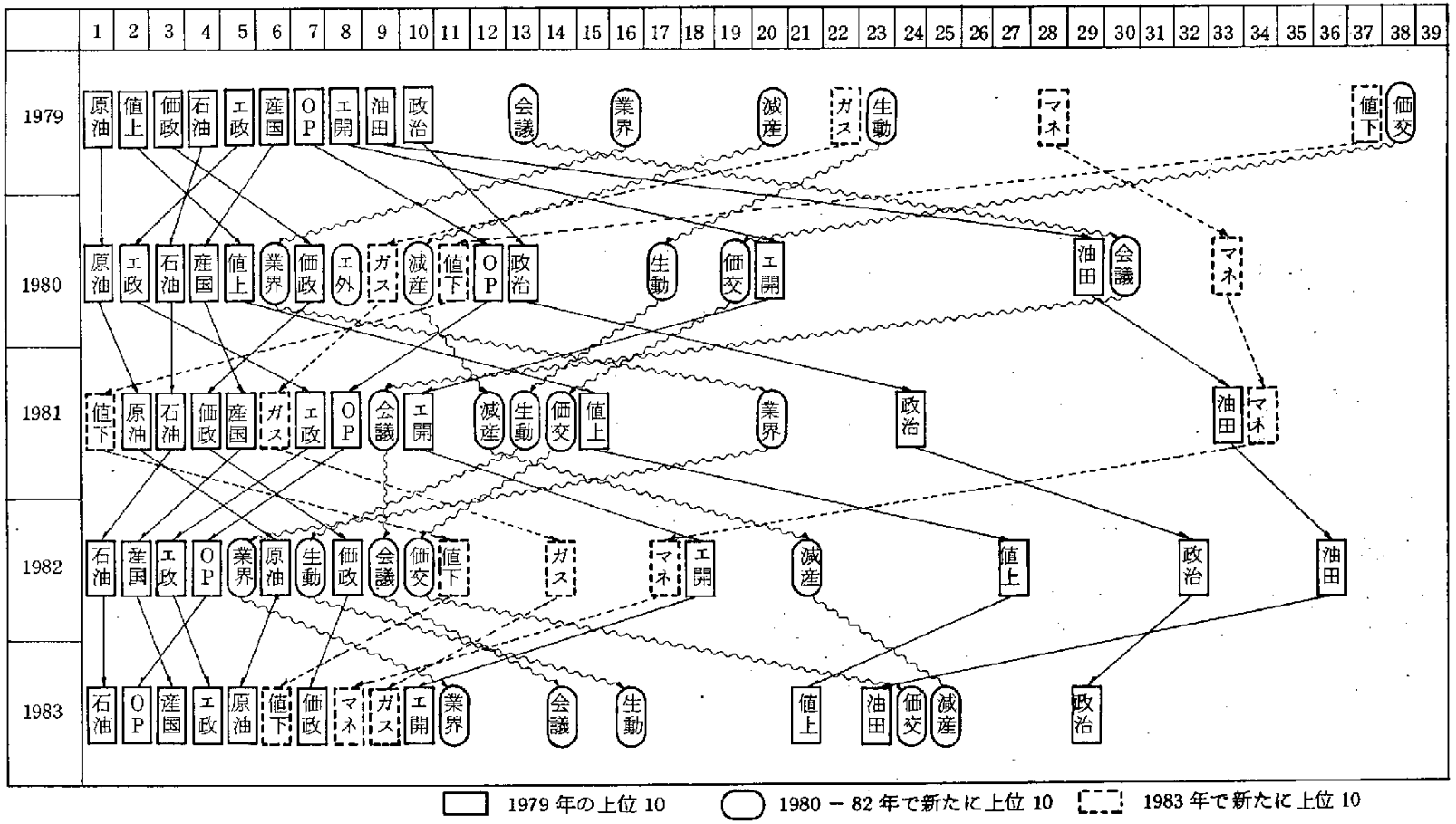


図 3.1－2　キーワード出現頻度順位変化(2)― ナイジェリア ―

共通に受けた時系列変化を示していると言うことができよう。すなわち，第２次石油危機後の原油価格の高騰を受けて，石油需要の減退と非ＯＰＥＣ原油の台頭により減産を余儀なくされ，価格設定と生産調整のための石油会議，石油業界・メジャーの圧力を経て８３年３月の原油価格の大幅値下げと追いつめられていく過程をキーワードの順位変化が示しているように見受けられる。

サウジアラビアの場合，この他では増産，生産計画というキーワードの変化が目立つが，これはスウィング・プロデューサとしてのサウジアラビアの特徴を示すものであろう。ナイジェリアの場合は，８２年，８３年のオイルマネーのキーワードが順位を上げていることと，油田あるいはＬＮＧプラントの開発に絡んでエネルギー開発のキーワードが上位を占めていることが大きな特徴となっている。

このようにキーワード出現頻度の順位変化を見ることによっても，ＯＰＥＣ加盟国の共通点，共通な時系列変化，固有の特徴を把握し，徴候的な変化を大づかみにとらえることはできそうである。

(3) 原油価格の変動と値上げ・値下げキーワードの変化

図3.1－1，図3.1－2に示したように，キーワードの出現頻度数と出現頻度順位の年次変化を分析した結果は，いくつかのキーワードがタイム・ラグをある程度持って増減していることを示唆するものである。このようなキーワード間のタイム・ラグを持った動きを過去何年間か把握しておいて近時点でキーワード出現頻度数の月次変化に過去とは異なる徴候が現われていないかを見出そうとすることはある程度有効な手法と考えられる。そこで，図3.1－3ではＯＰＥＣ加盟国において７９年と８３年で最も極端に相違した動きとなった値上げと値下げのキーワードの出現頻度数を原油価格の変化とともに示した。

値上げのキーワードは８０年末まで出現頻度数が多いが，８１年以降急速に減少し，値下げのキーワードの出現頻度数がこれに替って増

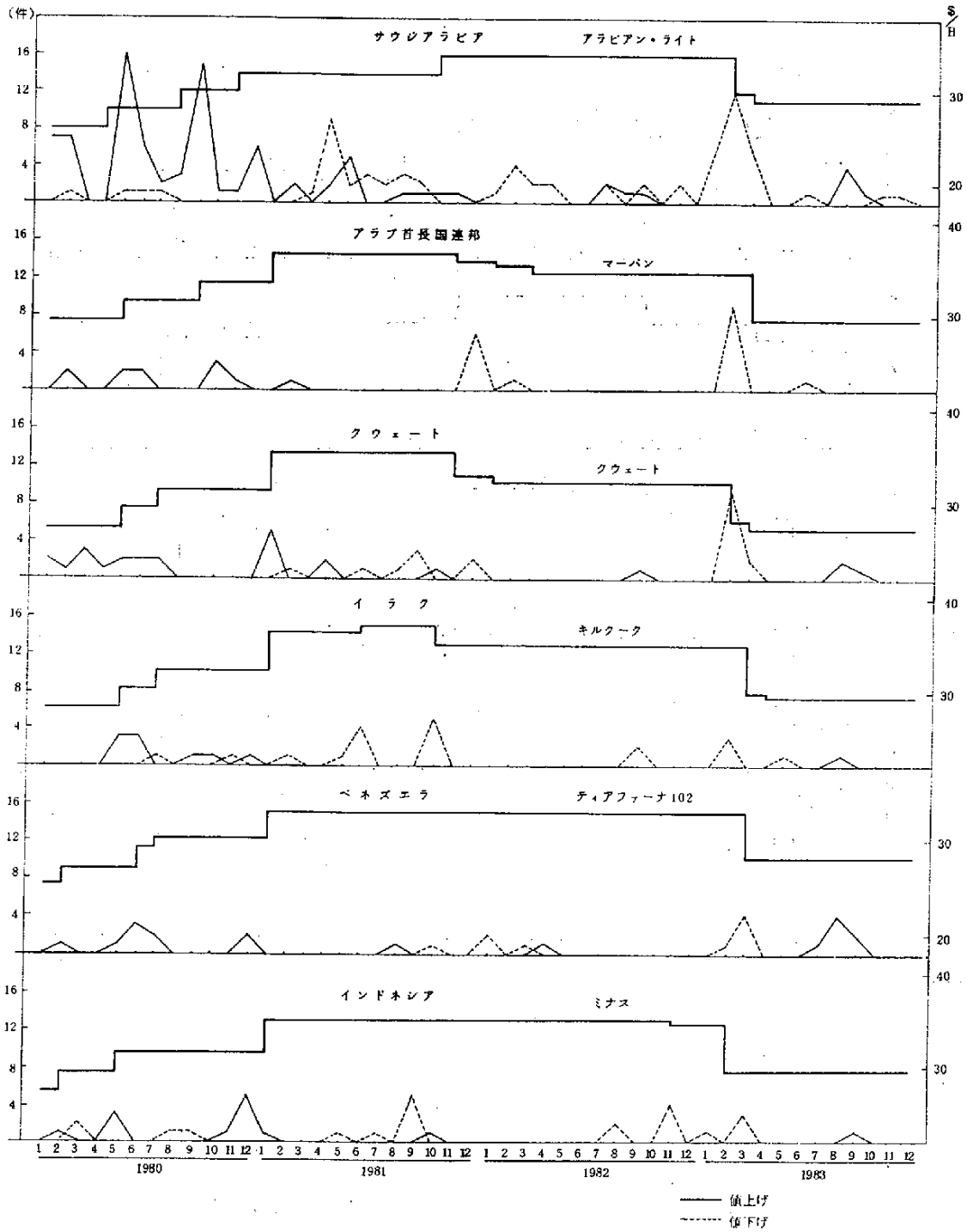


図 3.1－3　ＯＰＥＣ加盟国の原油価格と値上げ・値下げキーワード(1)

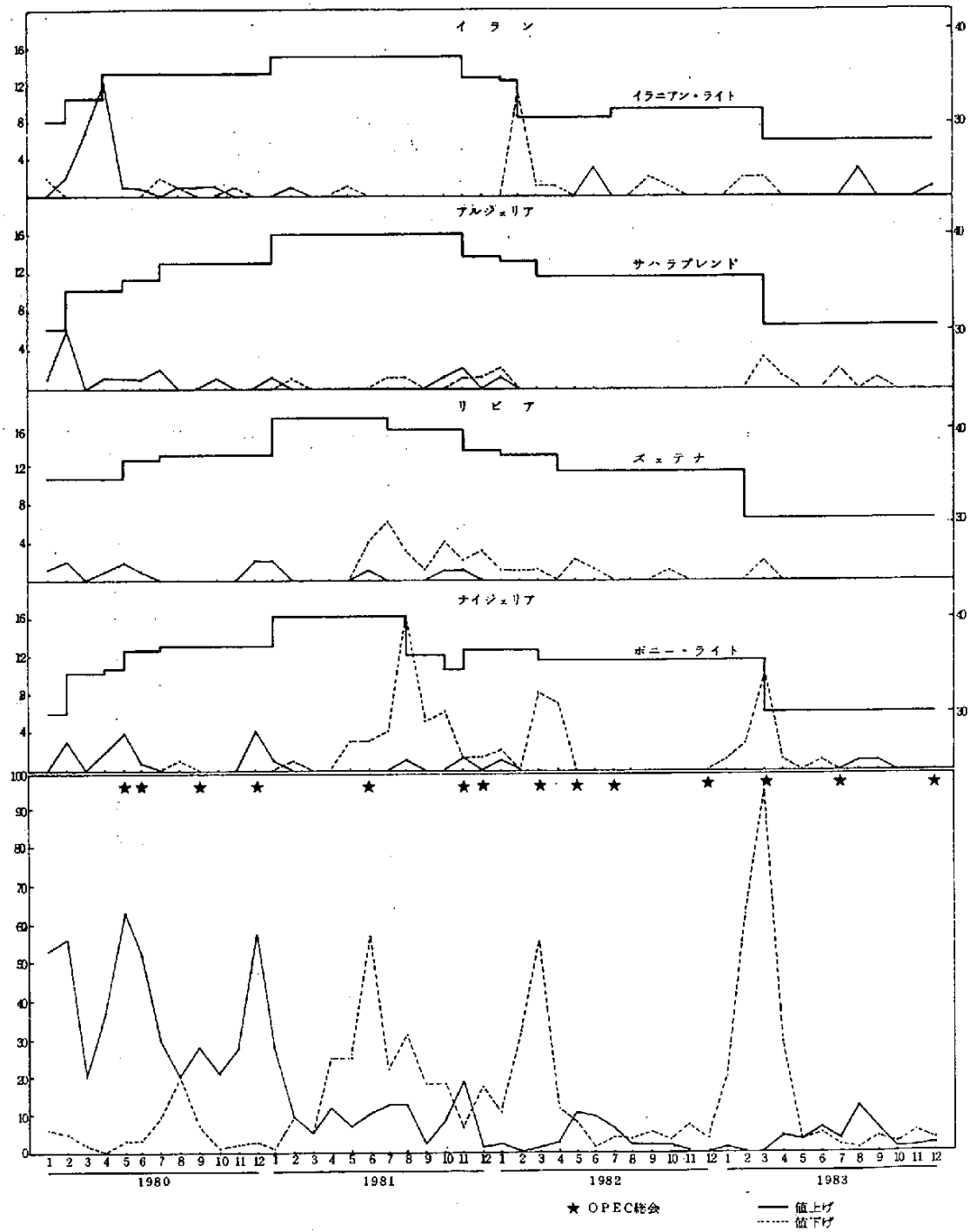


図3.1－3　OPEC加盟国の原油価格と値上げ・値下げキーワード(2)

大している。値上げ及び値下げのキーワード全体のピークの変化は，図3.1－3に示すように，ＯＰＥＣ総会の開催とかなり深く相関している。これに対して，ＯＰＥＣ加盟各国の値上げ，値下げキーワードの出現頻度数のピークは必ずしも全体のピークとは一致せずそれぞれの特徴を示している。

８０年における値上げは，イラン，アフリカ３カ国など急進派の動きにおされて，値上げを手控えていたサウジアラビアが値上げに踏み切ったのに対して他国が追随値上げをしていく様子がよく現われている。８１年の下半期はリビアの値下げに始って，ナイジェリアの大幅値下げ，サウジアラビアの値上げと他国の値下げによる価格体系の統一へと波及していく様子がよくわかる。その後は，値下げを手控えるサウジアラビアに対して，イラン，アフリカ３カ国を中心とした国々は値下げに走り，サウジアラビア主導で行われた ８３年３月の原油価格値下げに波及していく様子がよくうかがわれる。

キーワードの月次変化を追いかけることは，この例で示したように各国間での波及効果を分析するのに有効と考えられる。

(4) キーワード出現頻度数の月次変化

(3)では，原油価格の値上げ，値下げというある意味で両極に分離したキーワードの出現頻度数の月次変化を追いかけてみたが，ここでは，(1)，(2)で取り扱った年次変化の中でタイム・ラグを持って出現頻度数の増減がみられるキーワードを値上げと値下げのキーワードの間にはさんで月次変化を分析してみる。図3.1－4にサウジアラビアとナイジェリアの2カ国について，月次で国名キーワードの出現頻度数と各キーワードの出現頻度数の変化を示す。

サウジアラビアの場合は，値上げ，値下げのキーワードの他に，増産，エネルギー外交，エネルギー政策，減産，生産計画，業界，メジャーのキーワードを選択した。ナイジェリアの場合は，エネルギー政

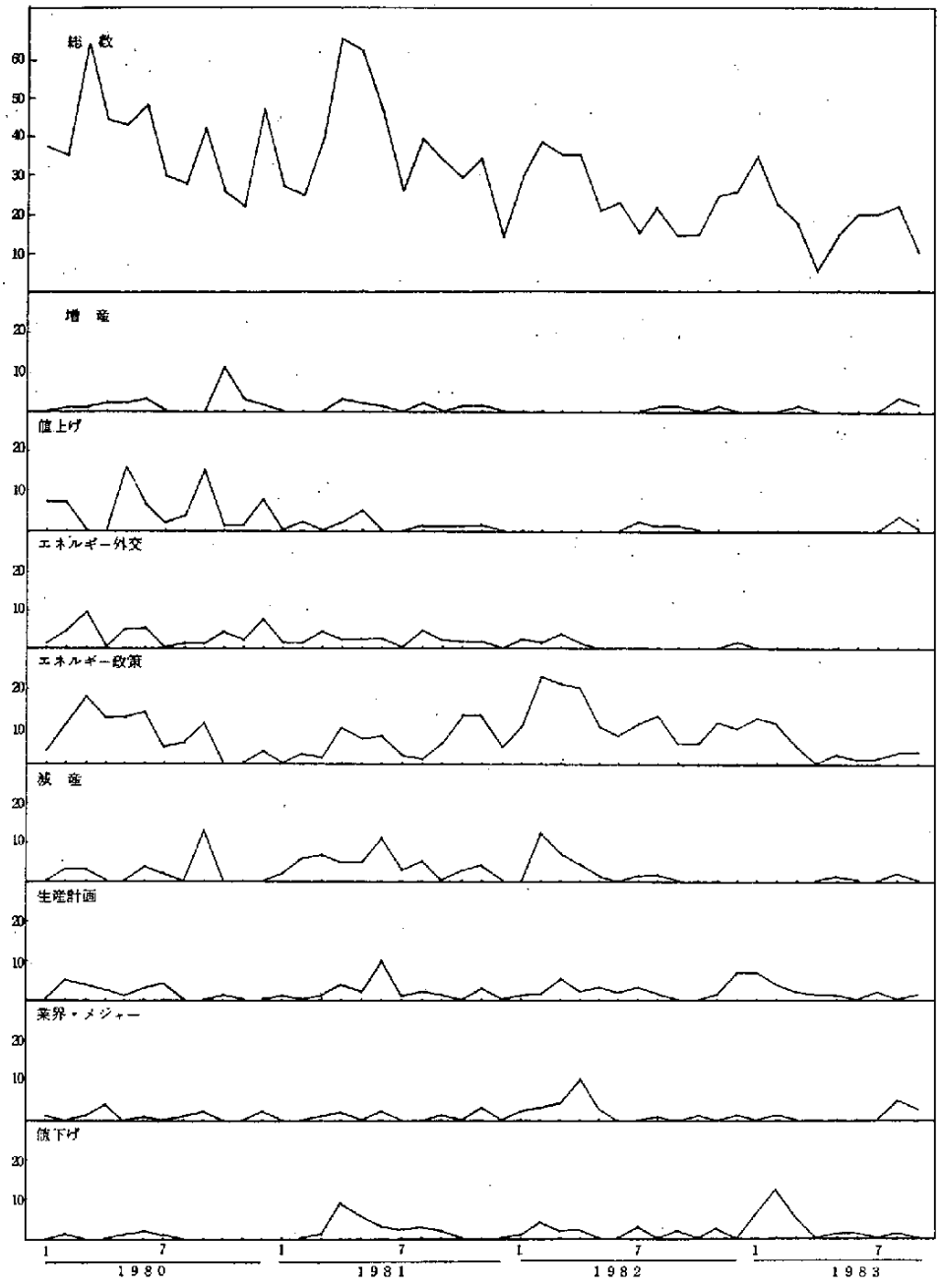


図 3.1－4　キーワード出現頻度数の月次変化(1)

— サウジアラビア —

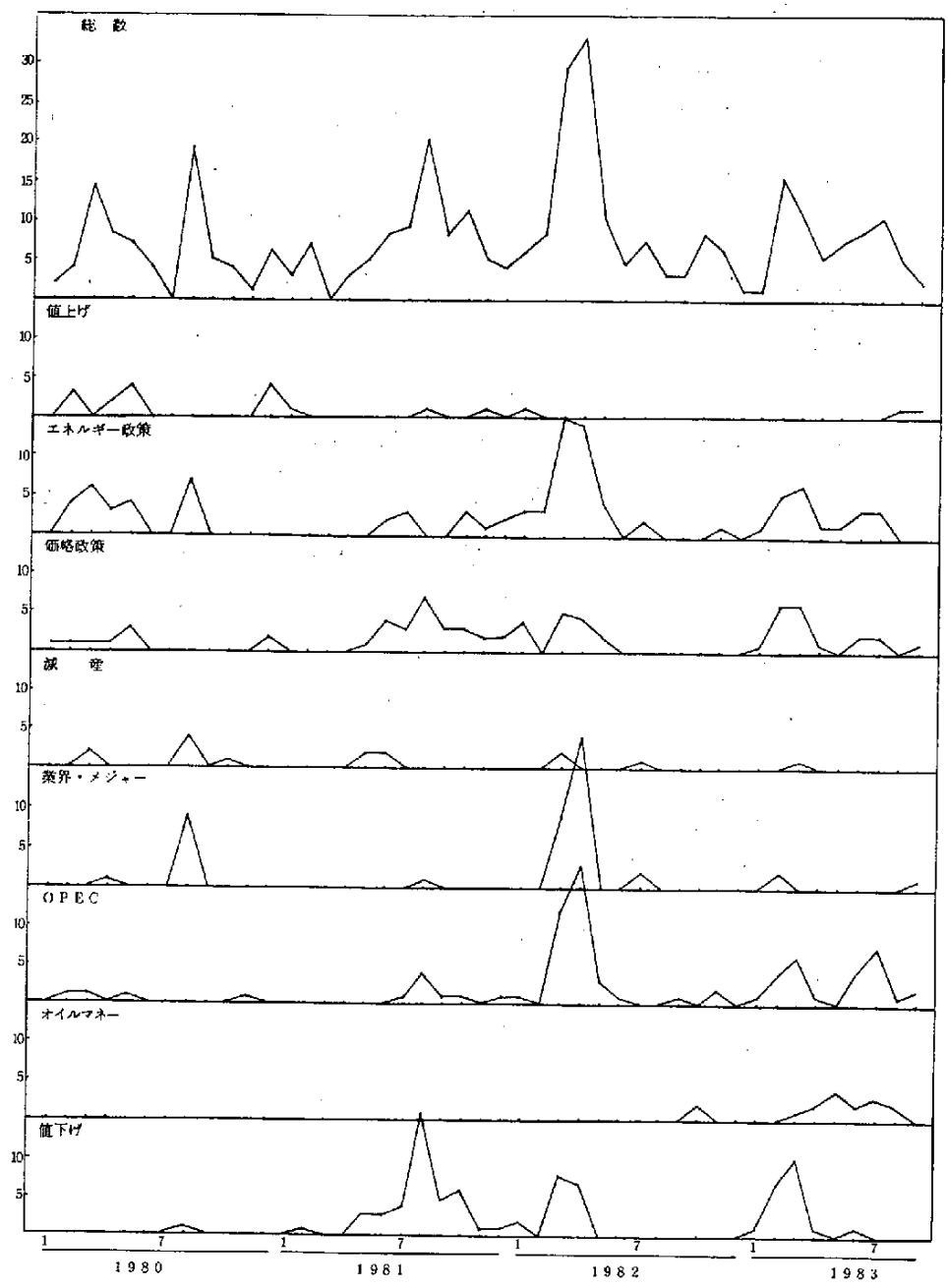


図 3.1－4　キーワード出現頻度数の月次変化(2)
— ナイジェリア —

策，価格政策，減産，業界・メジャー，ＯＰＥＣ，オイルマネーのキーワードを選択した。

月次変化の中で特徴的なことは，値上げを実施している８０年にサウジアラビアにしてもナイジェリアにしても減産の話題がすでにあがっていたことである。サウジアラビアの場合，８０年９月に減産の話題がかなり高まっているが，イラン・イラク戦争のため翌月は増産の話題に転じた。イラン・イラク戦争勃発による混乱が一段落ついた結果，８１年に入ってからは石油需要の減退と非ＯＰＥＣの台頭で減産を余儀なくされ，原油販売激減のあおりを最も大きく受けたナイジェリアは８１年８月に原油価格値下げに踏み切り増販を図らざるを得なくなったと考えられる。その後もＯＰＥＣ加盟国に対する減産圧力は続き，石油業界・メジャーの圧力も加わって財政逼迫のナイジェリアは８２年３，４月にも大きな圧力を加えられた。この時はサウジアラビアもナイジェリア援助に関して圧力を加えられ，エネルギー政策のキーワードが増加する形となっている。８２年のサウジアラビアは，エネルギー政策のキーワードの他，生産調整を示す生産計画のキーワードが増大している。ナイジェリアも財政逼迫に関連してオイルマネーのキーワードが増大している。

値上げと値下げのキーワードの増減の間では，このように各種のキーワードの増減の動きがあるわけであるが，過去数年間のキーワードの年次変化をおさえた上で，近時点のキーワード出現頻度数の月次変化を分析するという手法を採用すれば，過去とは異なった現象が起っており，詳しく調べてみようとするきっかけになる情報は，ある程度入手できると考えられる。

3.1.3　キーワード出現頻度数の国別比較分析

(1)　ＯＰＥＣ加盟国の国別比較

キーワード出現頻度数のデータは，時系列の分析に利用できるだけでなく，1時点を取って横並びに色々な国々の特徴を分析するためにも利用できると考えられる。図3.1－5と図3.1－6では，ＯＰＥＣの盟主国であるサウジアラビアを基準座標として，その他のＯＰＥＣ加盟国のキーワードの出現頻度状況を７９年と８３年の2時点を取って比較してみた。表3.1－2で示した各キーワードの出現頻度数を国名キーワードの出現頻度数で除して百分比としたものをデータとして用いており，サウジアラビアの百分比を大小順に並べて点線でプロットし，サウジアラビアのキーワード順に従って各国の出現頻度比を実線で示してある。

７９年の場合，各国独自の特徴がいくつか現われているが，サウジアラビアからのずれは８３年と比べれば小さいと言えよう。値上げと価格政策のキーワードが百分比でサウジアラビアよりも大きい値を他のＯＰＥＣ加盟国は共通に示している。増産と生産計画のキーワードの百分比は逆にサウジアムビアが他の国に比べて高い。７９年においてサウジアラビアが原油価格値上げに対して取った比較的穏健な行動と増産によって果したスウィング・プロデューサとしての役割が表現されていると考えられる。

イランはイラン革命を反映して，反政府運動，政治などのキーワードが高く出ており，他のＯＰＥＣとは大きく異ったパターンになっている。カタール，アルジェリア，インドネシアなどは天然ガスに関連してガスのキーワードの百分比が高くなっている。クウェート，ベネズエラ，インドネシアは，製油所，オイルサンド，油田などの開発に絡んでエネルギー開発のキーワードが高まっているのが特徴である。

８３年は，原油価格の値上げ，石油需要の減退，非ＯＰＥＣ原油の台頭，代替エネルギーの進展，石油供給過剰，石油業界・メジャーの

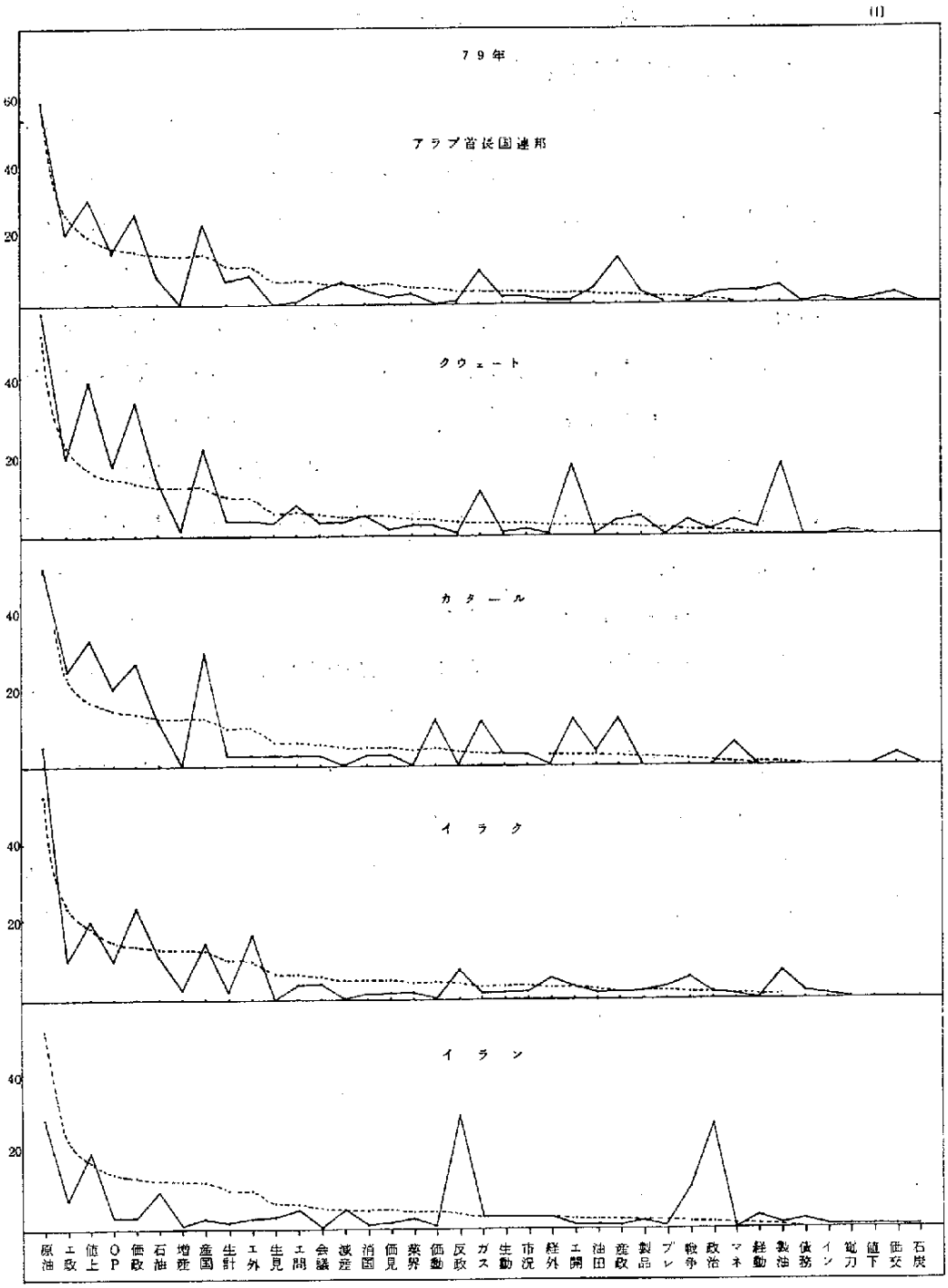
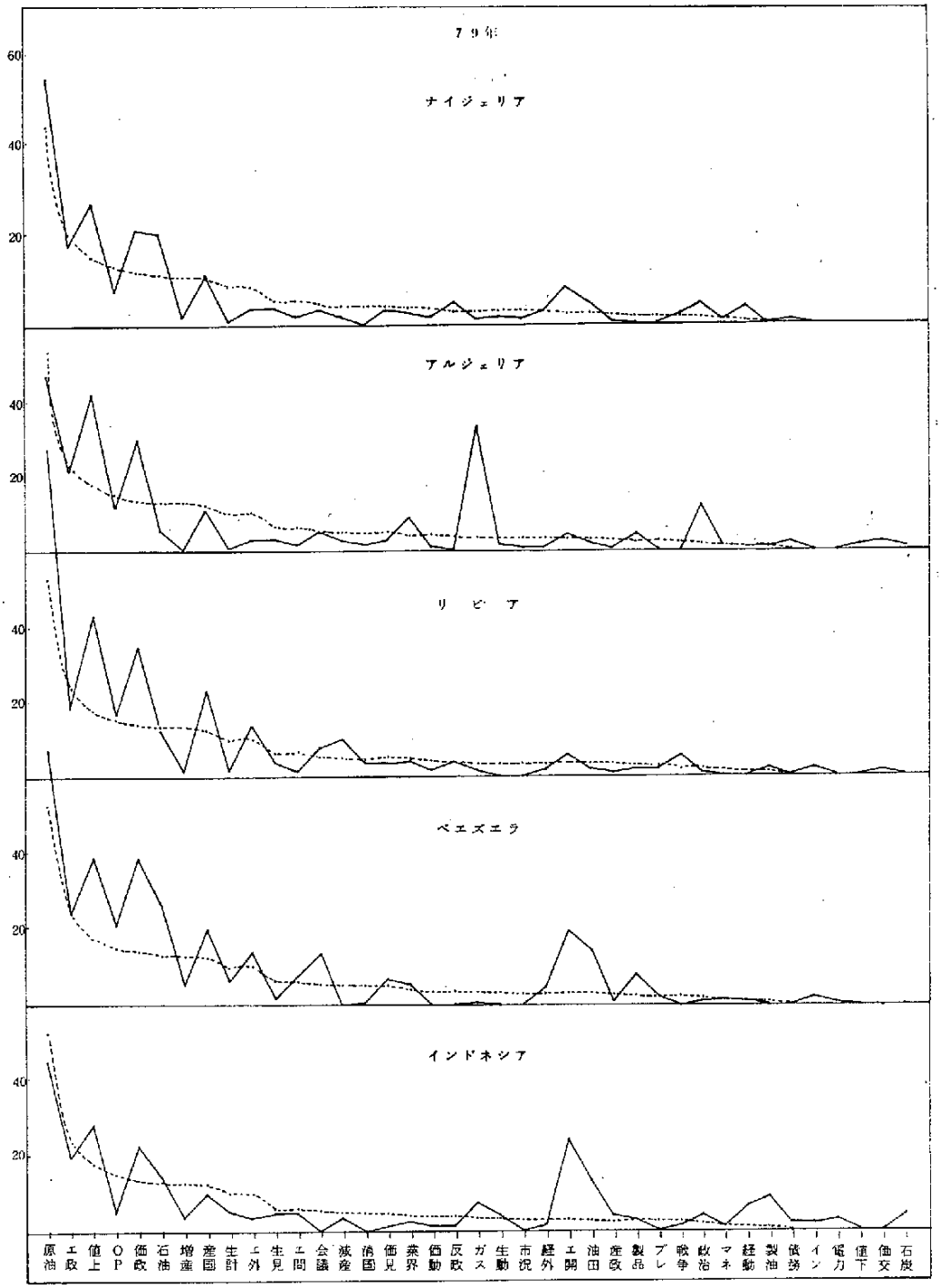


図3.1－5　サウジアラビアとその他ＯＰＥＣ加盟国との
キーワード出現状況の比較（'７９年）


79年
ナイジェリア
アルジェリア
リビア
ベエズエラ
インドネシア
60
40
20
原油
エ政
値上
OP
価政
石油
増産
産国
生計
エ外
生見
エ問
会議
減産
消国
価見
業界
価動
反政
ガス
生動
市況
経外
エ開
油田
産政
製品
プレ
戦争
政治
マネ
経動
製油
供給
イン
電力
重下
価交
石炭

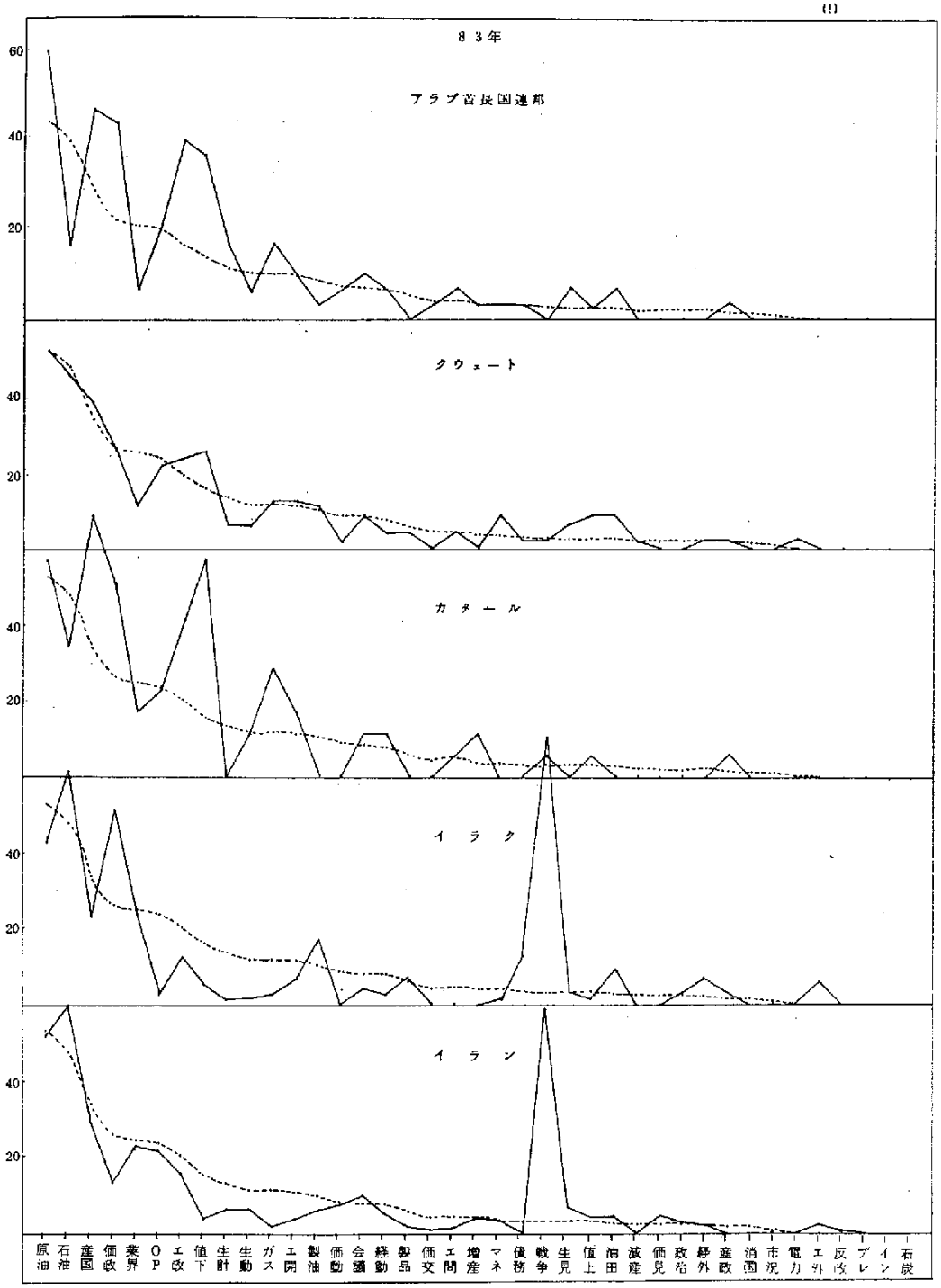
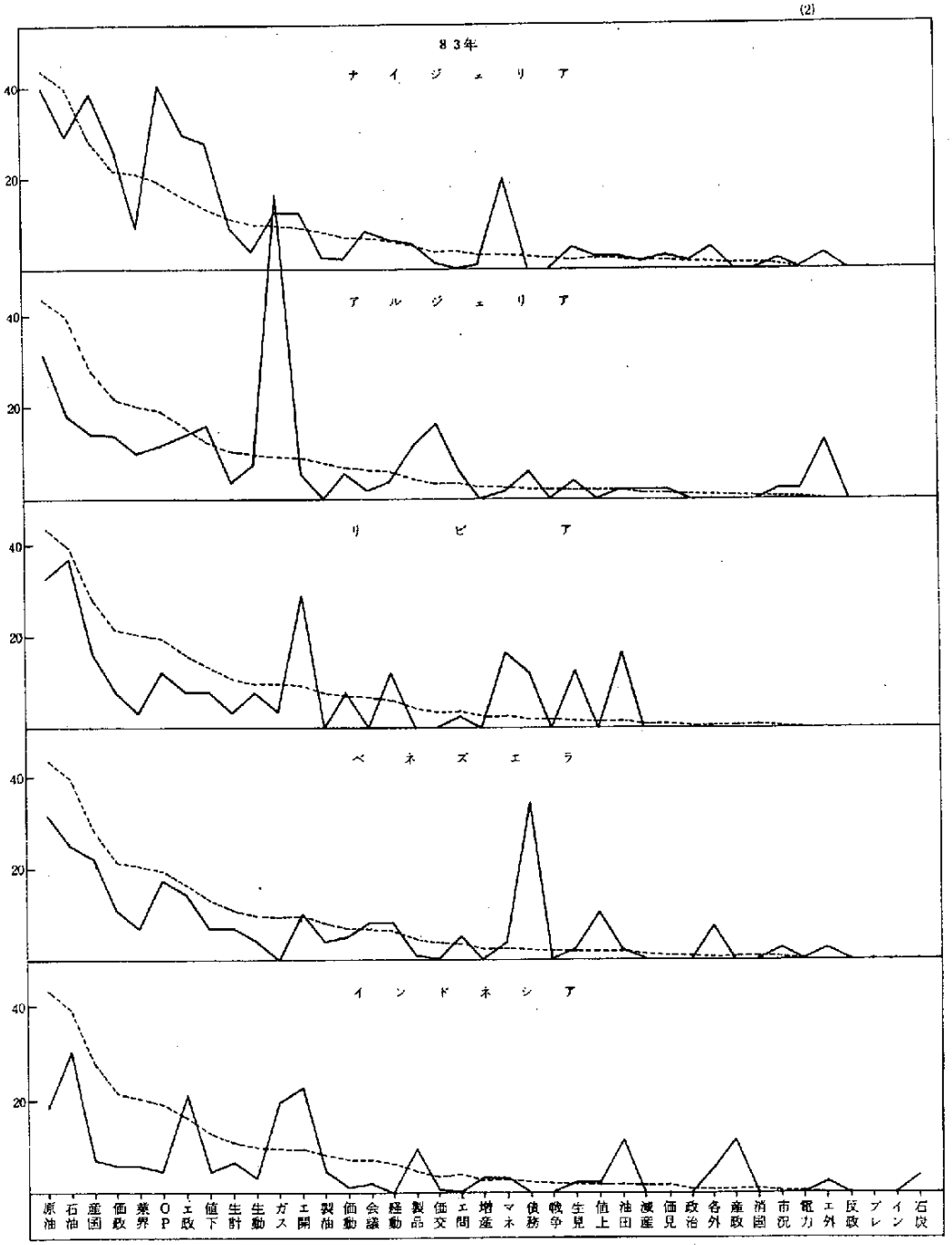


図3.1－6　サウジアラビアとその他ＯＰＥＣ加盟国とのキーワード出現状況の比較（'８３年）


83年
ナイジェリア
アルジェリア
リビア
ベネズエラ
インドネシア
40
20
原油
石油
産国
価政
業界
OP
エ政
値下
生計
生動
ガス
エ開
製油
価動
会議
経動
製品
価交
エ間
増産
マネ
債務
戦争
生見
値上
油田
減産
価見
政治
各外
産政
消国
市況
電力
エ外
反政
プレ
イン
石炭

圧力，原油価格の値下げと７９年からの大きな変動にもまれた結果，各国ともサウジアラビアの基準軸からはかなりずれた自国の特徴を出しているように見受けられる。イラン，イラクに関しては両国間の戦争の話題に関する百分比が高く，アルジェリア，インドネシア，カタールは７９年比べて一層天然ガスに関する話題が高まっている。ナイジェリア，リビア，ベネズエラに関しては，オイルマネーの減少と対外債務の話題が他の国に比べて強く出ているのが特徴である。

このように一つの国を基準軸としておくことにより，他の国との特徴の違いを分析していくことは，同じＯＰＥＣ加盟国であっても異なった性格を有することをクローズ・アップすることになるので有効な手法と考えられる。

(2) 先進国のエネルギー源別記事出現頻度の比較

ＯＰＥＣ加盟国が７９年から５年間でかなり大きな変化を受けたのと裏腹にして，先進国のエネルギー源に対する関心も大きく変化したと考えられる。図3.1－7では，米国，英国，西独，フランスの４カ国について，石油，石炭，ガス，電力，原子力，新エネの６種のエネルギーについて関連のキーワードで検索した記事インデックスを加算して総数を求め，各国名キーワードの記事総数で除した百分比に直してレーダー・チャートを作成した。

米国の場合，７９年は石油への関心が最も高く，他のエネルギー源はそれほどでもなかったが，８０年には石炭・新エネへの関心が高まり，８１年は石炭への関心が高まっている。８２年，８３年と石油供給過剰の影響を受け，８３年は新エネへの関心がぐっと低くなっている。原子力への関心は，７９年３月のスリーマイル島原発事故の発生により急速にさめている。

英国の場合，石油に対する関心のほか国内炭を産するので，石炭に対する関心が７９年も含めてもともと高い。８０年以降は，北海のガ

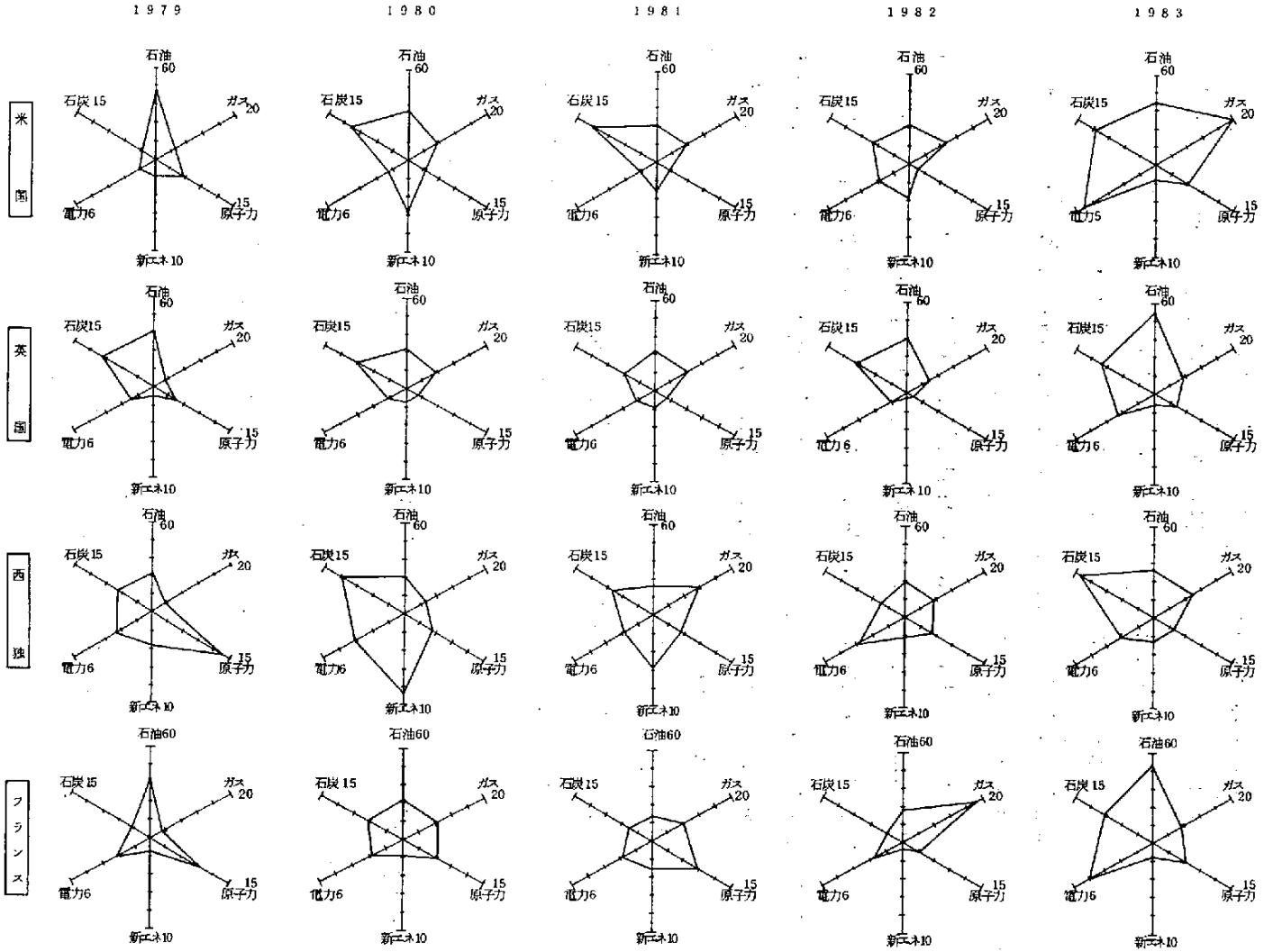


図 3.1－7　先進国のエネルギー源別記事出現の時系列変化

ス回収などガスに対する関心も高まっている。他の国と比べて新エネに関する話題が少ないのは大きな特徴である。

西独の場合，７９年は核燃料再処理などに関連して原子力の話題が多かったが，８０年は新エネ・石炭の話題が高まり，８１年はこれにガスが加わっている。原子力はスリーマイル島の事故の影響を受け，関心が低くなっている。８２年は石油供給過剰の影響を受け，比較的バランスのとれた関心度合となっている。

フランスの場合，７９年は石油と原子力による電力化計画に大きな関心が払われていたことがうかがえる。８０年には代替エネルギーとしてガス・石炭への関心も増大してきており，８１年は新エネに対する関心も伸びている。８２年は一転してソ連との天然ガスの話題が中心を占めている。原子力に対する関心度合が他の国に比べて高いのがフランスの特徴である。

この例では，エネルギー源に対する関心をレーダー・チャートで大づかみに見たが，各エネルギー源の中を各種のキーワードを駆使して同様な分析を行い，エネルギーに関する話題の細かい分析を行うことは可能と考えられる。

3.1.4　キーワード出現頻度数の主成分分析

(1)　時系列データの主成分分析

この分析実験では，７９年から８３年までの全データベース中のキーワードの出現頻度数を４半期別のデータとして収集し，キーワードを価格関連のキーワード，生産関連のキーワード，原油関連のキーワード，経済関連のキーワードの４つに分けて主成分分析を行った。図3.1－8にその結果を示す。

価格関連のキーワードの第１主成分は，図3.1－8に示すように，全体の３１％を説明するが，これは価格の値上げ，値下げに関連した

価格関連のキーワード

COMPONENT	EIGENVALUE	%	ACC.%		COPONENT 1	2	3	4
1	3.408	30.983	30.983	1価格交渉	−0.545	0.512	0.357	0.233
2	2.687	24.429	55.412	2価格据置	−0.1	0.008	0.916	−0.114
3	1.64	14.912	70.324	3価格政策	0.381	0.785	0.113	−0.395
4	1.07	9.727	80.051	4価格動向	−0.278	0.774	−0.055	0.204
5	0.763	6.94	86.991	5価格見通し	0.77	0.235	0.239	−0.354
6	0.61	5.547	92.538	6値 上 げ	0.93	0.192	−0.022	0.267
7	0.343	3.117	95.655	7値 下 げ	−0.59	0.595	−0.024	−0.309
8	0.251	2.278	97.934	8プレミア4販売	0.391	0.457	0.386	0.604
9	0.156	1.419	99.353	9価格管理	0.109	−0.498	0.592	−0.261
10	0.056	0.506	99.859	10価格カルテル	0.691	0.426	−0.306	−0.226
11	0.015	0.141	100	11価格戦略	0.657	−0.352	0.07	0.172

生産関連のキーワード

COMPONENT	EIGENVALUE	%	ACC.%		COPONENT 1	2	3	4
1	3.495	31.77	31.77	1生産管理	−0.218	−0.238	0.843	−0.065
2	2.004	18.22	49.99	2生産計画	0.503	0.617	0.042	−0.219
3	1.803	16.388	66.378	3生産動向	−0.482	−0.072	−0.145	0.726
4	1.303	11.841	78.219	4生産見通し	0.765	−0.314	0.409	0.064
5	0.737	6.696	84.915	5生産提携	0.694	0.571	0.184	0.073
6	0.505	4.591	89.506	6増 産	0.84	−0.381	0.072	−0.104
7	0.39	3.546	93.052	7減 産	0.525	−0.009	−0.173	0.744
8	0.302	2.745	95.797	8生 産 性	0.463	0.71	0.062	0.199
9	0.262	2.386	98.183	9操業開始	−0.083	−0.038	0.838	0.258
10	0.133	1.206	99.39	10操業再開	0.607	−0.613	−0.155	0.116
11	0.067	0.61	100	11操業停止	0.568	−0.332	−0.321	−0.173

原油関連のキーワード

COMPONENT	EIGENVALUE	%	ACC.%		COPONENT 1	2	3
1	4.797	39.971	39.971	1 O P E C	0.677	0.621	−0.01
2	3.154	26.284	66.255	2産 油 国	0.289	0.874	−0.066
3	1.023	8.524	74.779	3石 油	0.591	0.441	0.122
4	0.959	7.992	82.771	4原 油	0.908	−0.118	0.085
5	0.747	6.226	88.997	5製 油 所	−0.45	0.704	0.348
6	0.619	5.162	94.159	6油 田	0.802	−0.156	−0.098
7	0.272	2.267	96.427	7石油採鉱	−0.003	−0.664	−0.053
8	0.195	1.621	98.048	8石油業界	0.728	−0.359	0.382
9	0.126	1.046	99.094	9石油パイプライン	0.529	−0.211	0.713
10	0.049	0.406	99.5	10石油プラント	0.786	−0.048	−0.315
11	0.036	0.299	99.799	11石油会議	0.16	0.803	−0.05
12	0.024	0.2	99.999	12石油消費国	0.883	−0.118	−0.326

経済関連のキーワード

COMPONENT	EIGENVALUE	%	ACC.%		COPONENT 1	2	3	4
1	4.06	27.064	27.064	1対外債務	−0.153	0.696	−0.344	−0.447
2	3.9	25.999	53.062	2オイルマネー	0.228	−0.282	0.182	−0.746
3	1.73	11.531	64.593	3インフレ	0.93	0.064	−0.084	0.047
4	1.363	9.085	73.678	4経済成長	0.761	0.117	0.249	−0.005
5	1.067	7.112	80.79	5経済動向	−0.303	0.875	0.128	0.185
6	0.799	5.328	86.118	6経済見通し	0.295	0.761	−0.057	0.213
7	0.567	3.783	89.901	7経済外交	0.519	−0.463	0.348	−0.295
8	0.515	3.432	93.333	8経済協力	0.502	−0.226	−0.66	−0.009
9	0.35	2.334	95.667	9経済計画	0.284	−0.45	−0.63	−0.002
10	0.249	1.658	97.325	10経済政策	0.817	0.181	0.302	0.139
11	0.175	1.17	98.494	11国 有 化	−0.043	−0.526	−0.239	0.498
12	0.128	0.851	99.346	12借 款	0.039	0.709	−0.576	−0.269
13	0.056	0.389	99.734	13財政政策	0.267	0.746	0.174	0.178
14	0.026	0.175	99.91	14金融政策	0.561	0.364	0.039	−0.152
15	0.013	0.09	99.999	15産業政策	0.819	−0.072	−0.156	0.207

図3.1－8　時系列データによる主成分分析

軸であることを因子負荷量の情報は示している。第２主成分は価格政策・価格動向に絡んだ軸，第３主成分は価格据え置きの軸，第４主成分はプレミアム販売の軸と，第４主成分までで全体の８０％を説明している。

生産関連のキーワードの第１主成分は，生産計画，生産見通し，増産，減産のコンビで今後の生産の方向性に関する話題の軸で，３２％を説明している。第２主成分は生産性に関する軸，第３主成分は生産管理に関する軸，第４主成分は生産動向と減産に関する軸で，第４主成分までで７０％を説明している。

原油関連のキーワードの場合，第１主成分は原油，石油，ＯＰＥＣ，油田，石油業界，石油プラント，石油消費国がコンビとなっており，原油に関連する話題が増大する軸と考えられる。第１主成分は全体の４０％を説明している。第２主成分は，ＯＰＥＣ，産油国，石油会議で産油国の話し合いを示す軸となっている。第３主成分は石油パイプラインの軸である。第３主成分までで全体の７５％を説明している。

経済関連のキーワードの第１主成分は，インフレ，経済政策，産業政策のコンビでインフレ調整を含む政策絡みの軸となっている。第２主成分は経済動向・経済見通しの軸，第３主成分は，経済協力・経済計画・借款のコンビで対外援助の軸，第４主成分はオイルマネーの軸となっている。第４主成分までで全体の７４％が説明されている。

この主成分分析によって得られた軸は，５年間の時系列データの共通項として得られたものである。記事インデックスに付されたキーワードはおのおのが独立ではなく相互に依存しているので，主成分分析を通して中心軸となるキーワードを見通してから細かい頻度数分析に入っていく手法を採ることが妥当であろう。

(2)　クロスデータの主成分分析

この主成分分析の実験では，3.1.2の(1)で述べたＯＰＥＣ加盟国に

		固有値				因子負荷量			
	COMPONENT	EIGENVALUE	%	ACC.%		1	2	3.	4
	1	9.138	83.075	83.075	1 サウジアラビア	0.928	0.058	−0.23	−0.145
	2	0.863	7.847	90.922	2 UAE	0.977	−0.083	0.014	−0.037
	3	0.334	3.036	93.958	3 クウェート	0.957	−0.109	0.033	0.051
	4	0.215	1.95	95.909	4 カタール	0.932	−0.195	0.075	−0.121
	5	0.146	1.331	97.239	5 イラク	0.946	0.13	−0.201	−0.061
1979年	6	0.109	0.99	98.23	6 イラン	0.475	0.868	0.12	0.032
	7	0.074	0.676	98.906	7 アルジェリア	0.865	−0.096	0.433	−0.159
	8	0.053	0.479	99.385	8 ナイジェリア	0.976	0.034	−0.052	0.075
	9	0.033	0.302	99.687	9 リビア	0.976	0.026	−0.092	−0.087
	10	0.02	0.183	99.87	10 ベネズエラ	0.957	−0.109	−0.101	0.114
	11	0.014	0.129	100.	11 インドネシア	0.921	−0.102	0.101	0.345
	1	6.698	60.895	60.895	1 サウジアラビア	0.96	−0.123	−0.044	−0.045
	2	1.438	13.072	73.968	2 UAE	0.898	−0.193	0.104	−0.179
	3	1.012	9.204	83.171	3 クウェート	0.891	−0.079	0.054	0.122
	4	0.83	7.543	90.714	4 カタール	0.454	0.01	−0.548	0.694
	5	0.348	3.162	93.876	5 イラク	0.435	0.853	0.217	0.155
1980年	6	0.302	2.744	96.62	6 イラン	0.723	0.644	0.21	−0.047
	7	0.178	1.619	98.239	7 アルジェリア	0.146	−0.414	0.768	0.458
	8	0.089	0.811	99.05	8 ナイジェリア	0.89	−0.142	−0.045	−0.162
	9	0.06	0.543	99.594	9 リビア	0.902	−0.035	−0.101	0.025
	10	0.032	0.289	99.883	10 ベネズエラ	0.944	−0.132	−0.045	−0.189
	11	0.013	0.117	100.	11 インドネシア	0.862	−0.159	−0.041	−0.023
	1	6.629	60.264	60.264	1 サウジアラビア	0.922	−0.138	0.216	0.051
	2	1.617	14.702	74.966	2 UAE	0.821	0.14	0.373	0.107
	3	1.098	9.979	84.945	3 クウェート	0.864	0.048	0.195	−0.242
	4	0.501	4.557	89.502	4 カタール	0.753	0.413	−0.098	−0.386
	5	0.383	3.482	92.984	5 イラク	0.671	−0.6	−0.373	0.004
1981年	6	0.352	3.201	96.185	6 イラン	0.643	−0.56	−0.47	−0.138
	7	0.189	1.718	97.903	7 アルジェリア	0.236	0.704	−0.588	0.15
	8	0.091	0.829	98.731	8 ナイジェリア	0.832	0.105	0.089	0.408
	9	0.062	0.563	99.294	9 リビア	0.921	−0.207	−0.07	0.228
	10	0.05	0.454	99.748	10 ベネズエラ	0.87	0.044	0.3	−0.139
	11	0.028	0.252	100.	11 インドネシア	0.759	0.425	−0.236	−0.017
	1	7.415	67.41	67.41	1 サウジアラビア	0.938	−0.072	−0.171	−0.202
	2	1.314	11.948	79.358	2 UAE	0.909	−0.07	−0.145	0.238
	3	0.781	7.1	86.458	3 クウェート	0.942	−0.066	−0.121	−0.189
	4	0.687	6.248	92.706	4 カタール	0.885	−0.174	−0.065	−0.06
	5	0.327	2.976	95.681	5 イラク	0.597	0.775	0.175	0.03
1982年	6	0.193	1.757	97.438	6 イラン	0.718	0.648	0.205	−0.055
	7	0.126	1.15	98.588	7 アルジェリア	0.444	−0.424	0.784	−0.062
	8	0.065	0.587	99.175	8 ナイジェリア	0.922	−0.166	0.002	−0.227
	9	0.053	0.486	99.661	9 リビア	0.932	−0.018	−0.027	0.03
	10	0.026	0.235	99.896	10 ベネズエラ	0.88	−0.164	−0.153	−0.052
	11	0.011	0.104	100.	11 インドネシア	0.688	−0.115	0.03	0.698
	1	7.222	65.655	65.655	1 サウジアラビア	0.956	−0.006	−0.056	−0.037
	2	1.242	11.293	76.948	2 UAE	0.86	−0.248	−0.245	−0.126
	3	0.802	7.294	84.242	3 クウェート	0.965	−0.007	−0.058	0.03
	4	0.623	5.66	89.902	4 カタール	0.848	−0.279	−0.169	−0.242
	5	0.357	3.244	93.146	5 イラク	0.688	0.596	0.198	−0.275
1983年	6	0.244	2.22	95.366	6 イラン	0.773	0.527	0.127	−0.238
	7	0.181	1.645	97.011	7 アルジェリア	0.545	−0.571	0.475	−0.174
	8	0.146	1.33	98.341	8 ナイジェリア	0.878	−0.229	−0.245	−0.039
	9	0.094	0.853	99.194	9 リビア	0.809	0.245	0.114	0.409
	10	0.058	0.526	99.72	10 ベネズエラ	0.768	0.091	−0.343	0.337
	11	0.031	0.28	100.	11 インドネシア	0.731	−0.153	0.486	0.318

図3.1－9　クロスデータによる主成分分析固有値と因子負荷量

関する様々なキーワード（表3.1－2参照）の出現頻度百分比をベースデータとして，年次別に話題によるＯＰＥＣの結び付きを分析した。結果を図3.1－9に示す。

いずれの年次も第1主成分ではＯＰＥＣ加盟国が＋でリニア・コンビネーションしている。第1主成分はＯＰＥＣ加盟国に関する共通話題の軸となっている。７９年は第1主成分が８３％を占めており，乱している第2主成分と第3主成分は革命の起ったイランの話題とアルジェリアの天然ガスの話題である。８０年，８１年とは第1主成分による説明が６０％強へと低下した。これは，イラン・イラク戦争によるイラン・イラク両国の分離と天然ガスに関するアルジェリア，カタールなどの話題が強まったためである。８２年，８３年には石油供給過剰の影響を受け，戦争の話題もいくぶんマンネリ化して，共通の話題である第1主成分が若干とり戻した。

このように，クロスデータによる主成分分析も全体を大づかみする上で有効な情報を与えることができると考えられる。

3.1.5　キーワード出現頻度数のクラスター分析

この分析実験においても，3.1.2の(1)で述べたＯＰＥＣ加盟各国に関するキーワード（表3.1－2参照）の出現頻度の百分比データを使用した。クラスター分析では，類似度の距離を設定する方法として最短距離法，最長距離法，メジアン法，重心法など様々な方法を採ることができるが，図3.1－10では記事出現にみるＯＰＥＣ加盟国の類似度をクラスター分析の最長距離法で年別に分析した結果を示している。国名はペルシャ湾岸6カ国，アフリカ3カ国，その他2カ国の順に並べて樹状図をクラスター化の順に従って描くことにより結果の表示を行っている。

７９年は，イランを除いて湾岸の4カ国が比較的早くサウジアラビアのラインに合流しているが，８０年，８１年，８２年とイラクは戦争の

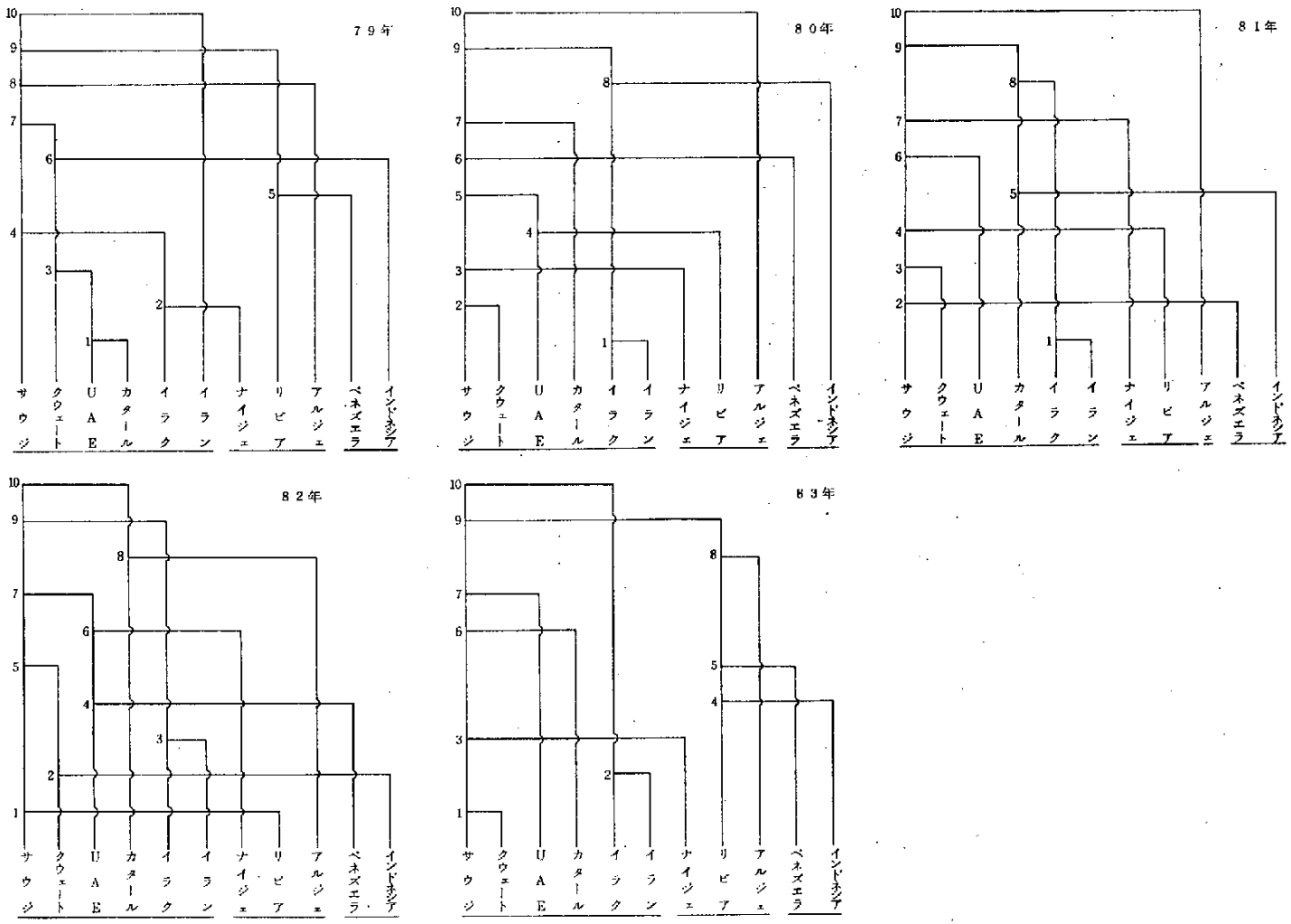


図 3.1－10 記事出現にみるＯＰＥＣ加盟国の類似度 ― クラスター分析（最長距離法） ―

ため，カタールは天然ガスへの力点の移行のため，サウジアラビアのラインからはある程度離れ，ナイジェリア，リビア，ベネズエラなどの方に近くなっている。８３年は，石油供給過剰の影響を受け，戦争を継続しているイラン，イラクを除いては，７９年の話題の類似度によるパターンに近くなってきているようである。

クラスター分析も，相互に依存しているキーワードを類似度ごとに分類し，カテゴリー化するのにかなり有効な手法と考えられる。クラスター分析によりキーワードをカテゴリー群に分類し，主成分分析により分類したキーワードの中から主流となる軸を見い出すという形で分析を進め，見い出した主成分の軸に沿った観点からキーワードの出現頻度を細かく分析していけば，種々の話題の変化と特徴に関する有効な分析が可能になると考えられる。

3.1.6　キーワード出現頻度数の回帰分析

この分析実験では，ＯＰＥＣ加盟国の国名キーワードの出現頻度数と図3.1－11の下にリストアップした様々なキーワードの出現頻度数をデータベース全体から４半期データとして作成し，単純相関分析で国名キーワードとの相関関係を分析している。分析結果は，サウジアラビア，アラブ首長国連邦，クウェート，カタールの４カ国について図3.1－11に示した。

サウジアラビアの場合，価格政策，値上げ，プレミアム販売，減産，エネルギー外交，ＯＰＥＣ，原油，経済動向といったキーワードとの相関係数が高くなっている。この中では，プレミアム販売との相関係数が高いことだけは直接の結びつきで説明できない要素となっている。

アラブ首長国連邦の場合，政治関連のキーワードや操業再開のキーワードとの相関が高まっているが，これはアラブ首長国連邦の国名キーワード出現のパターンが比較的イランの国名キーワード出現のパターンと

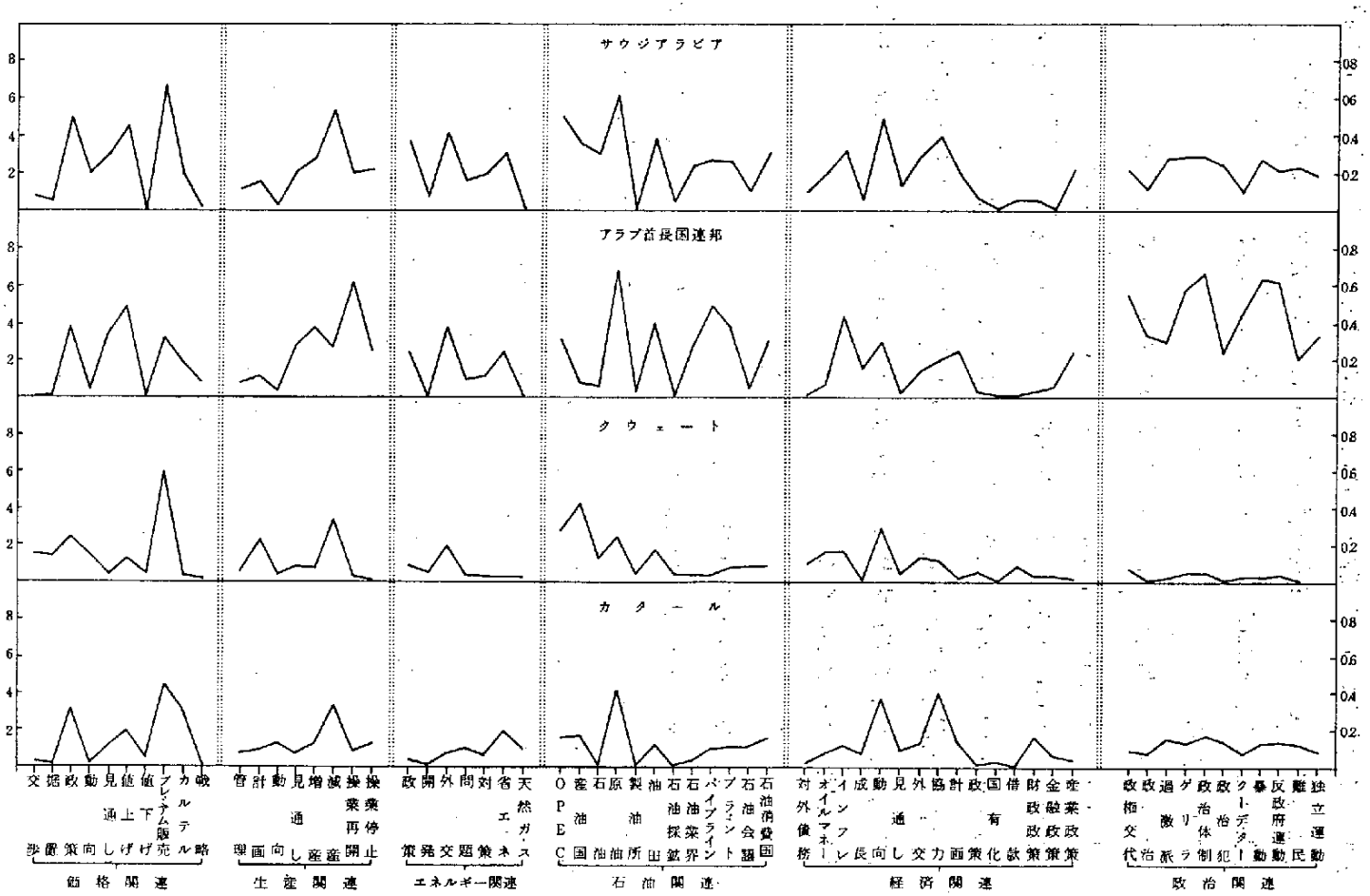


図3.1－11　OPEC加盟国キーワード出現頻度数と各種キーワード出現頻度数の時系列相関

類似していたということであり，直接関係はない。むしろイラン革命の勃発に関連してアラブ首長国連邦も話題にのぼったが，最初の興奮が落ち着いてくるとともに国として取り上げられる話題も下火になったということであろう。

クウェートに関してプレミアム販売の相関が高いのは，サウジアラビアとは異なって問題となった当事国であるので妥当な結果と言えよう。

このように回帰分析は，主成分分析やクラスター分析に比べると今一つ効果が低いと見受けられがちである。しかし，主成分分析，クラスター分析，出現頻度の時系列分析，比較分析などを通して関係の有無を十分に把握した上で使用すれば，有効な回帰分析の結果と予測式を得られる可能性もあると思われる。

3.1.7　定量化分析の機能と課題

以上の世界エネルギー情報によるキーワードの出現頻度数分析の結果に基づくと，文章情報データベース・システムにおける定量化分析の機能としては，表3.1－3に示したような機能が必要と考えられる。

頻度数生データとしては，時系列頻度数，1時点の共出現頻度数，特定の集合における全キーワードの頻度数が必要と考えられるので，データベースから一定のルールで効率よく検索する機能を設定する必要がある。

頻度数生データは，百分比などに加工する必要があり，加工を効率よく行うためには頻度数マトリックスの行あるいは列単位での加工ができることが望ましい。また，別々に検索してできた頻度数データのマトリックスを結合したり，絞り込んで抽出する機能も必要である。こうした機能を合わせて頻度数データ加工機能を設定する必要がある。

検索結果を常に主記憶に有することはおそらくできないと考えられるので頻度数生データあるいは加工データを保存するための機能の設定も

表3.1－3　定量化分析システムの機能

サブシステム名	機能名	機能内容
頻度検索	1. 時系列頻度数	・検索対象となるファイルは，オリジナルなデータベースの他，検索機能で絞り込んだ2次保存ファイルも含む。 ・時系列の期種は月次・四半期・年次・期間・期種と指定して時系列のキーワード出現頻度数を検索する。
	2. クロス頻度数	・検索対象となるファイルにおけるキーワードの共出現頻度数を検索する。
	3. 全キーワード頻度数	・検索機能で絞り込んだ2次保存ファイルを対象として含まれる全キーワードと頻度数をリストアップする。 ・2つの保存ファイルを比較して新しいキーワードの出現，頻度数の変化を分析する。
頻度数データ加工	1. 計算機能	・伸び率，構成比，頻度数間の四則演算をマトリックスの行・列単位で行い，加工データを作成する。
	2. マトリックス部分抽出	・分析用にオリジナルのマトリックスから任意の行あるいは列を抽出し，部分マトリックスを作成する。
	3. マトリックス結合	・異なった検索で得られた頻度数マトリックスから，行あるいは列の内容に注意して結合マトリックスを作成する。
頻度数データ保存	1. データ保存	・分析用の頻度数生データあるいは加工データを作業用の中間ファイルあるいは保存用ファイルに蓄積する。
定量化分析	1. 主成分分析（次元縮小）	・時系列あるいはクロスの出現頻度数に基づいて，キーワード間の独立な主成分を求め，キーワードの出現頻度数情報をより少ない成分で説明する。
	2. クラスター分析（分類）	・クロスの出現頻度数に基づいて，各種キーワードをカテゴリー化，分類する。
	3. 回帰分析（相関関係・予測）	・時系列あるいはクロスの出現頻度数分析に基づいてキーワード間の相関関係を分析する。
グラフ出力		・キーワードの出現頻度数の生データ，加工データを折れ線グラフ，棒グラフ，レーダーチャートなどで出力する。

必要である。

本体である定量化分析の機能としては，分類のためのクラスター分析，次元縮小のための主成分分析，相関関係，予測のための回帰分析の機能を設定する必要があると考えられる。

こうした分析機能に基づいて分析の主軸と対象となるキーワードを決め，過去のキーワードの年次変化を把握しながら近時点の月次変化を追い，変化の兆しとなるような徴候的な情報を見出すことが必要となる。そのためにはビジュアルな形でグラフ出力をできるだけ豊富に行う機能が必要と考えられる。

ところで，本研究の結果，記事インデックスに付したキーワードは，必ずしも定量化分析に向いたものではないことが判明した。定量化分析のためには，＋－の方向性と事実か観測かという識別を持ったキーワードを付けることが必要と考えられる。もちろんこれで十分なキーワードになるわけではないが，こうしたキーワードに基づいて出現頻度の定量化分析を行えば，かなりの情報が整理されてくると考えられる。今後のデータ整備の課題である。

3.2　キーワード自動抽出システムの現状

ベタ書きの漢字かな混じり文章からキーワードを自動抽出する試みは，研究レベルから一部，実用化の段階に入っている。日本科学技術情報センター（ＪＩＣＳＴ）では，数年前から，雑誌論文のタイトルから自動抽出したキーワードを補助キーワードとして検索の用に供している。また，データベースのディストリビュータ，コンピュータ機器メーカ，計算センターなどがキーワード自動抽出機能を搭載した，文章情報の蓄積・検索システムを提供する動きも出始めている。

実用化システム，あるいは実用化に近いレベルにあるシステムといっても，現状では精度100%の完全自動化は無理で，最終的には人間の判断によるボ

スト・エディティング（後処理）に委ねているものが多い。文章情報総合解析システムが目指す水準には，まだかなりの距離があるのが現状である。

それでも，キーワード自動抽出を使用して文章情報データベースや索引誌などの作成を試みる動きが強まっていることは大変歓迎すべきことである。文章情報総合解析システムの開発には，研究室レベルでの先端的な実験，研究は無論必要だが，実用的な技術を着実に積み上げていく努力も欠かせない。

キーワードは何よりも情報検索の場で，利用者が求める情報を的確に探し出す役割を負っている。データベースの作成にあたる人々の間では，自動抽出されたキーワードの品質に不安を抱く人も少なくない。現状ではポスト・エディティングという，いわば人間と機械の分業ないしは共同作業を前提にした半自動化システムであるのはやむを得ない。

たとえ半自動化システムであっても，これによるデータベース作成のコスト減とスピード化は，総合利用推進に寄与するところは大きい。現在の技術レベルでは，一足飛びに完全自動化を求めることは無理であり，多少の未熟さに目をつぶっても実用化が可能な技術は積極的に開発を進め，データベース利用の現場で磨き上げられていくべきであろう。そこから得られたデータ

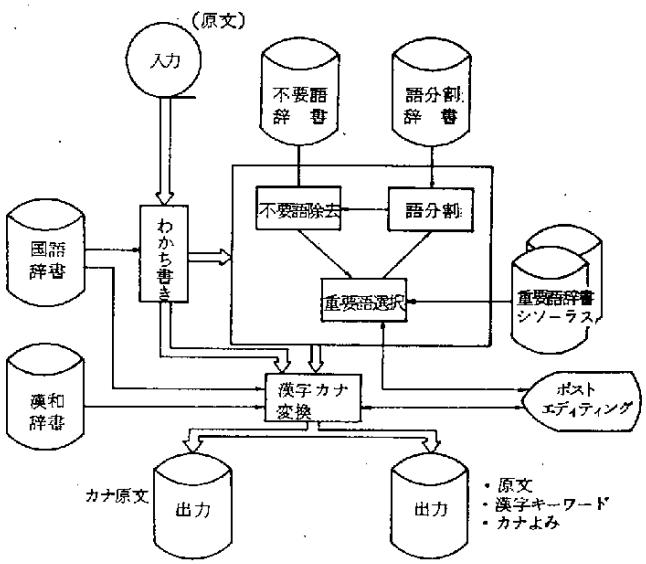


図3.2－1　キーワード自動抽出システム概念図

や経験や知識が，総合利用システムの開発に反映されることが期待できる。

本節では，こうした観点から，実用的なキーワード自動抽出システムの開発のために，現在の技術で達成可能な機能および手法を探ってみた。とくに将来，文章情報利用面で大きな比重を占めると思われる新聞記事情報を中心に考察した。

3.2.1 キーワード付与の条件

従来から人手で行われているキーワード付与が，どのような仕組みになっているか，そこにどんな問題があり，自動抽出を採用することでどんな影響を受けるか，などを考察することは自動抽出システムの実用化を図るうえで，必要なことである。

人手によるキーワード付与は，おおむね次のような過程を経る。

① 主題の分析

② 主題の抽出

③ 索引の選択

主題（ Subject ）とは，索引づけの対象となる文章を，他の文章と識別しうる重要な概念を指す。主題分析，主題抽出は索引者が文章を読み，その文章を理解して，抽出すべき主題を認識する過程である。主題分析を客観的に行う方法として，換言すれば索引づけの整合性・画一性を保つために，通常主題が存在すると思われるカテゴリーをあらかじめ定めておき，このカテゴリーに存する概念は一応主題である。と見なして抽出する方法をとることが多い。設定されるカテゴリーは，索引づけシステムが取り扱う分野が科学技術文献であるか，経済・ビジネス情報であるか，法律・判例の類であるかによって各々異なるのは当然である。たとえば日本経済新聞社のＮＥＥＤＳ－ＩＲは，経済に関する新聞・雑誌記事検索サービスであるが，ここでは抽出すべき主題のカテゴリー，言い換えれば記事に付与されるべきキーワードのカテゴリーとして，次

の７種類を設けている。

・会社キーワード：国内外の企業名

・団体キーワード：国内外の官公庁，各種団体，機関名

・人名キーワード：国内外の人名

・品目キーワード：商品・サービス・システム・工法の名称

・業界キーワード：業界の名称

・項目キーワード：政治・経済・経営・社会における様々な事象を表わす用語

・地域キーワード：国内は県別，海外は国別の地域名および海洋名

索引者が主題を認識・抽出して，それを表現するために選ばれた用語がキーワード（またはディスクリプタ）である。キーワードの選択に関して，従来から統制語方式をとるかフリータームにするかの問題があることはよく知られている。この両者は一長一短であり，併用方式が採用されることが多くなっている。併用方式といっても，統制語の語数が非常に少ないフリーターム中心型と，統制語で表現しきれない特殊な概念の場合にのみフリータームを使用する型とがある。

フリータームを使用する場合，主題に対してどこまでSpecific なキーワードを与えるかが問題になる。「農産物輸入に関する日米貿易交渉」について書かれた記事に対し，「日米農産物輸入交渉」とするか，「農産物輸入」と「日米交渉」という２つ（または「日米貿易」を加えて３つ）のキーワードを与えるか，あるいは「農産物」「輸入」「日米」「貿易」「交渉」といった単語レベルにまで分解するかは，一定の基準がない限り索引者によって異なるだろう。この問題は，キーワード自動抽出において，複合語・合成語をどの程度にまで分割するかに係る問題でもあるが，日本語は分かち書きする習慣がないためか，単語という観念が薄く，情報検索の利用者にとってはある程度の長さ以下であれば，複合語の方が受け入れられやすいようだ。

キーワード付与にあたっては，検索時における再現率の向上に，最も注意が払われる。情報検索の利用者にとっては，ノイズは容易に見分けがつくが，検索もれを発見することは通常不可能である。検索もれが生じるのは，索引者と利用者との間で，主題に対するキーワードの選定に食い違いが生じるからであり，当然ながらフリータームでは起こりやすい。長所より欠点の方が多いにもかかわらず，統制語方式がデータベース作成側にとって捨て難い最大の理由がここにある。

キーワード自動抽出システムの設計にあたっては，情報検索の実際面で生じている問題について，十分な配慮が払われなければならない。

3.2.2 キーワード自動抽出の手法

冒頭に述べたように，現在実用化されつつあるキーワード自動抽出システムは，人間によるポスト・エディティングを前提にしている。自動抽出システムに求められる機能を区分けすると，

① 単語や用語の抽出

② 漢字・カナ変換

③ キーワードの選定

になると考えられる（図3.2－1）。現在の技術レベルでは，①および②についてかなりの精度で実現できるが，③の機能を満足させる普遍的な手法やアルゴリズムの確立はまだまだ難しく，最終的には人間の判断に頼らざるを得ない。しかし人間が文章を読んでキーワードの付与・入力を行うやり方に比べると，データベース作成のコスト面でも迅速さの面でも格段の違いがあることは明白である。

以下に，各機能についての最近の動向や手法について考察する。

(1) 単語・用語の抽出

普通に書かれた現代文から，単語や用語を分離する分かち書きは，ほぼ実用の域に達しているとみてよい。よく知られているように分か

ち書きの手法としては，

- 字種の変化に着目する
- 単語辞書や国語辞書によるマッチング
- 語の結合規則を利用する

などがあるが，字種の変化と辞書とのマッチングを組み合わせると，かなりの成果が期待できる。キーワード候補となる用語を抽出するだけなら，字種の変化に着目するとともに，活用語の語尾変化，ひらがなの自立語，数詞等の小規模な辞書を使うだけでもある程度可能である。しかし，品詞の判定，助詞の種類によって用語が文中でどんな役割を果しているかの推定，複合語の分割などを行うには国語辞書が必要である。

用語の抽出では，抽出の単位として複合語はそのまま抽出するか，単語に分割するかが問題となる。単語分割の方が国語辞典のエントリーも少なく処理も容易であろう。文章中や複合語の中での位置情報を保持しておけば，検索時に複合語でアクセスするのに障害はないとの見方も成り立つ。

しかし，複合語によって表現されている特定の概念は単語分割によって全く失なわれてしまう。企業名などの固有名詞や「ロッキード事件」「日米貿易摩擦」のような時事用語を分割することは意味がないし，このような用語でなくても複合語はやはりそのままで抽出されるべきであろう。キーワード抽出における分かち書きは，単なる単語の認定でなく，重要語の選択のための処理であり，重要語としての特殊な意味の重さは，複合語の中により多く見受けられるからである。

ただ，新聞記事などによく見られるケースであるが，テニヲハが省略されたことによって，本来，複合語でないものが連続した漢字やカナの文字列として抽出されるのは避けなければならないのは当然である。このような文字列を単語分割でなく，意味の切れ目で分割するこ

とは難しい問題ではあるが，語と語の結合規則を利用した分割処理の研究開発に力を注ぐ必要があろう。

また，後述するように重要語の選定のための頻度分析では，単語レベルでの頻度を調べることも必要になるので，単語分割機能も合わせて備えておくべきである。

(2) 漢字カナ変換

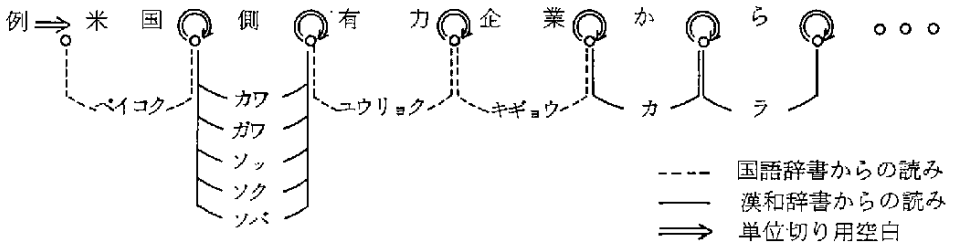
日本語文章情報データベースを検索する際には，キーワードはカタカナで入力する必要がある。また索引誌などの作成の際にも見出し語の配列にカナよみが使用される。

最近のワードプロセッサやパーソナル・コンピュータの普及ぶりからみて，将来は端末側でカナ漢字変換を行い，データベースには漢字キーワードでアクセスするので，キーワードのカナ読みは必要がなくなるとの見方もあろう。ただ，ユーザー・オリエンテッドな立場から見ると，カナ漢字変換のような端末操作上の負担をユーザーに強いることへの疑問も強い。データベースの分野・性格からみて，一義的に決定できる漢字キーワードでも，端末側だけの処理では，数多くの同音語の中から選ばなければならないようなケースも考えられ，漢字・カナの変換は，やはりデータベース・システムが負担すべきであろう。また，文章情報総合利用の一つの形態として点字訳や音声合成による朗読（音訳）などへの利用を考えると，漢字カナ変換の技術は欠かせないものがある。

漢字キーワードにカナ読みを与えるには，単語辞書とのマッチングによるのが一般的である。マッチングがとれない漢字については，一定の規則によるカナ振りが試みられる。最も単純な方法は，その漢字の前後が平仮名なら訓読み，そうでなければ音読みにする方法である。しかし，音には漢音，呉音，唐音などの違いがあり，また前後の漢字の読み方によっては濁音化，半濁音化，拗音化する。さらに「山桜」

「足音」などのように漢字１字の自立語（和語の名詞）から成る熟語もある。従って，漢字が持つ読みの数と種類，活用語になるか否か，送り仮名が付けられるか省略されるか，１字の自立語になるかなどの種別や濁音化・拗音化する場合の前後の読みなどを持つ漢字辞書を作成して，少しでも精度を上げる必要がある。

日本アイ・ビー・エム㈱では，漢字カナ混じり文の単位切りとカナ変換を同時に行うため動的計画法（Dynamic Programing）を用いたシステムを開発している。このシステムでは知識ベースとして，国語辞書，「確率付き」漢和辞書（１つの漢字が持つ複数の読みとその読みが出現する確率を持つ）および「３－gram」と呼ぶ一種の辞書を使用している。「３－gram」とは，カナ２文字のあとに，どんなカナがくるか，起り得るあらゆる可能性についての確率を，それぞれのカナ３文字に与えたものである。入力された文章中の漢字文字列に対して，国語辞書にある２～３字の文字列にはその読み方を与え，そうでない漢字には漢和辞書から，１文字ずつすべての読みと確率を与える。この様にして，カナ読みの有向グラフを作成して（図3.2－2），Dynamic Programing法により，出現の最大確率になるようなカナ列のルートを探索する。またカナ３－gram のデータからの確率や，音



（情報処理学会第２７回全国大会講演論文集）

図3.2－2　入力例と有向グラフ

訓別，用語の活用語，語尾変化などによる出現確率を加味して，読みと切れ目とを推定する。このシステムは文章情報の自動点訳のために開発されたものであるが，キーワードの自動抽出への応用も考えられている。

漢字カナ変換で最も厄介な問題は固有名詞のカナ読みである。固有名詞の読み方には全く規則性がないから，企業名，団体，機関名，人名，地名などをできるだけ多く辞書に登録するほかに手はない。この辞書に登録される固有名詞の数は，膨大な数にのぼることが予想されるが，固有名詞辞書は漢字カナ変換と同時に，キーワードの選定にも有効な働きをすると考えられる。

(3) キーワードの選定

キーワードの選定にあたっては，まず不要語（ストップワード）の除去が必要である。不要語としては以下のものが考えられる。

(イ) 付属語や特定の品詞。当面は名詞（「引き上げ」「見通し」など動詞の連用形の複合語や名詞と動詞の複合語を含む）以外の品詞は不要語とする。

(ロ) 接頭語や接尾語

(ハ) 数値を含む語，数量情報。数量情報はキーワードには成り得ないが，文章情報の利用において重要な情報であり，別の配慮を要する。

(ニ) 一般語。ただし単語としては不要語となるが，複合語の中ではキーワードになり得るものが多いので選択がむずかしい。

以上の処理を経た用語（キーワード候補語）の中から，キーワードを選定する技術が自動抽出の中で最も困難な課題である。

重要語の選定には，固有名詞などを収録した重要語辞書あるいはキーワード辞書とのマッチングと出現頻度などによる重要度の分析・評価という2つの方法を組み合わせて行うのが最も現実的である。このうち，どちらが欠けてもキーワードの選定はうまく機能しないと思わ

れる。

まず，文章中に現われる固有名詞はすべてキーワードである，と考えてよいだろう。固有名詞の判定には辞書を用いるしかない。企業・団体・機関をはじめ人名，地名などできる限り多くの固有名詞を収録した辞書が必要である。人名をすべて辞書に登録するのは不可能だが，語型のパターン解析や姓・名の部分一致を使えば，辞書にない人名もある程度カバーできるだろう。企業・団体の略称，「中曽根首相」などのような姓と肩書きによる表記についても対応措置が欠かせない。

時事用語やそれ以外の重要語，キーワードとなり得る用語も辞書とのマッチングによって抽出する。単純な方法だが，新聞記事などでは固有名詞や重要語は繰り返して使用されることが少ないので，やむを得ない。このような辞書は，あらかじめ用意しておくことも可能だが，むしろデータベースの作成を重ねることで蓄積されていくものである。

辞書マッチングだけでは，キーワードの選択は限界があるから頻度分析による重要度の評価を行ってキーワードを決定する。頻度分析では単語レベルに分割して集計する方が良い結果が得られるかも知れない。「このシステムは」とか「同計画が」などの表現はよく用いられるので，頻度の高い単語を含む複合語は，それ自体の頻度は低くても重要度が高いと推定される。

重要度の評価では，単なる出現頻度だけでなく，文章の長さ，その語が出現した位置などの諸要素を織り込んだ係数化が必要だろう。他の文章と比較して，その文章にだけ特徴的に現われる用語を探り出す方法もある。１つの文章だけでなく，文章情報データベース全体での用語の頻度を集計する。この用語がそれぞれの文章に均一に出現すると仮定すれば，当該文章における頻度は，全体の文字量と当該文章の文字量の比に比例する。このようにして求められた理論的な頻度と実際の出現頻度を比較して，実際の頻度の値が十分大きければ，その

用語は当該文章に特徴的に表われる語だと判定できる。この方法で問題になるのは，常に全体との比較が必要なことで，データベースに新たにデータが追加されるたびに全体の頻度と文字数を更新する必要がある。またすべてのキーワード候補語を常時抱えていなければならない。しかし，一つの文章の中で，パラグラフ単位に特徴的に表われる用語の認定に役立つ。

漢字を含むキーワード候補語に対して，特定の分野で出現頻度の高い漢字に着目するのも一つの方法であろう。たとえば，エネルギーの分野では，「燃」「油」などの文字の出現頻度が，他の分野に比べて非常に高いとすると，これらの文字を含む用語を重要語とすることができるだろう。

頻度分析では，単純な方法から高度の手法まで様々な方法が試みられるべきであろう。対象となる文章が記事であるか論文であるかの違いや，分野の違いによって，もっとも適合した方法が選べるようなオプションが望ましい。

シソーラスがあれば，選定されたキーワードから対象分野に適合しない語をふるいにかけることが可能になる。また統制語が必要な場合にも，シソーラスを使用して，キーワードの追加を行う。これによって追加された統制語は，文章中から抽出された語と厳密に区別されねばならないが，人間が付与した統制語に比べてバラつきがないだけ使い易いかも知れない。

3.2.3 新聞記事を対象とした自動抽出

新聞記事，特にニュース記事は比較的キーワード自動抽出に適合したタイプの文章情報である。その理由としては，新聞によって多少の違いはあっても文章のスタイルや用語・漢字表記・仮名づかいなどが統一されていることや，主題を表わす用語が記事中に明示的に表現されている

ことなどである。特に新聞記事のスタイルが，まず重要な事柄を先に書いていくという，いわゆる“逆三角形”の構成になっていて，記事のリードや前半の部分に重要語が集中していることが大きな利点である。これによって，キーワードの選定において，キーワード候補語が記事のどこに現われたかの位置情報を生かすことができるし，記事全文でなくリードの部分だけを対象に自動抽出を行っても，キーワードを取り出すことが可能である。

その一例として，ＮＥＥＤＳ－ＩＲの記事要旨（記事の見出しにあたるもの）と記事抄録（記事の前半の４００字以内の部分をとり出したもの）から，キーワード候補語を抽出した例を図3.2－3に示す。用語の抽出には日外アソシエーツ㈱が作成した「用語管理プログラム」を使用している。ＮＥＥＤＳ－ＩＲではキーワードとして統制語を主体とした消印方式をとっているが，これらを登録した辞書によってキーワード候補語を選定した結果もあわせて示してある。キーワードがフリータームであれば，それに対応する統制語を追加している。キーワードとして選定された用語の中で「親」の文字が印された語が追加キーワードである。例えば，エネルギー見通しの記事では，原油価格というフリータームから，原油と価格動向の２つの統制語が追加されている。因に人手によって付与されたキーワードは，エネルギー見通しの記事では，日本エネルギー経済研究所，エネルギー，エネルギー開発，需給見通し，民間統計である。また日韓国鉄の記事には，国鉄，韓国国鉄，鉄道旅客，オリンピック，日韓関係，販売戦略，外人観光客，韓国，陸運政策，対外関係の各キーワードが付与されている。

このシステムでは，連続した漢字・カナ文字列は分割しないで抽出する方法をとっている。このため，エネルギーのような最重要語がキーワードになっていない。この他にも選にもれた用語の中にはキーワードとして望ましい用語がいくつかある。複合語の中からキーワードを分離

今後のエネルギー見通し、石油需要は来年から一転して増加傾向――エネ経済研発表。

日本エネルギー経済研究所（生田豊朗理事長）は二日、最近のエネルギー情勢と今後の見通しを発表した。それによると、（1）世界の石油需要の減退は八三年で底を打ち、八四年から増加に転じる（2）このため、ＯＰＥＣ（石油輸出国機構）、非ＯＰＥＣともに来年から生産量が拡大する（3）原油価格は現在なお高すぎるが、来年いっぱい凍え据かれる――などで、中東地域での突発事態を除けば大きな変化はないとしている。

エネルギー見通し
石油需要
一転
エネ経済研発表
日本エネルギー経済研究所
生田豊朗理事長
エネルギー情勢
世界
減退
打ち
ＯＰＥＣ
石油輸出国機構
非ＯＰＥＣ
生産量
原油価格
据え置
中東地域
突発事態

キーワード候補語

キーワード
〔団〕 日本エネルギー経済研究所
〔団〕 ＯＰＥＣ
〔品〕 規 原油
〔事〕 規 価格動向
〔地〕 世界
〔特〕 原油価格

日韓の国鉄が共通パス――ソウル五輪に照準、外人客誘致へ来年発売めざす。

日本と韓国の国鉄が協力、主として外国人向けに両国の国鉄を共通利用できる「ジャパン・コリア・レール・パス」（仮称）を発売しようという計画を検討していることが明らかになった。韓国は一九八八年のソウルオリンピック開催へ向けて外国人観光客誘致に一段と力を入れるが、日本側も韓国経由の観光客が訪れてくれれば売り上げ増につながると積極的な姿勢を見せている。すでに実務家同士が数回協議しており、共同研究チームを発足させ、早ければ来年中にも発売にこぎつけたいという。日韓の国鉄が営業分野でタイアップするのはこれが初めてで、関係者は大いに期待している。

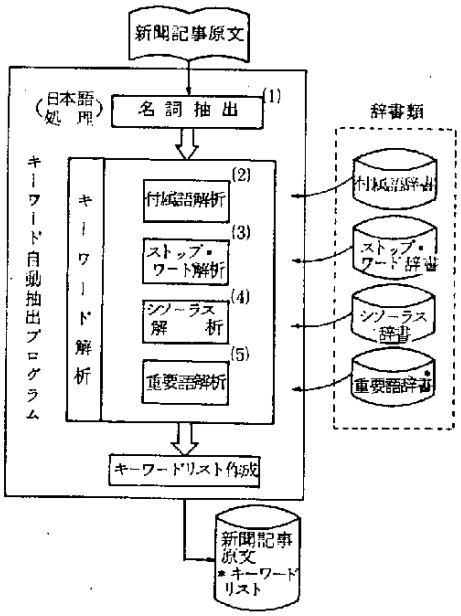
日韓
国鉄
共通パス
ソウル五輪
照準
外人客誘致
発売
日本
韓国
外国人向
両国
共通利用
ジャパン・コリア・レール・パス
仮称
計画
ソウルオリンピック開催
外国人観光客誘致
入れ
日本側
韓国経由
観光客
売り上げ増
実務家
数回協議
共同研究チーム
営業分野
タイアップ
初め

〔団〕 国鉄
〔品〕 規 旅行観光
〔項〕 規 オリンピック
〔地〕 韓国
〔地〕 日本
〔特〕 観光客
〔特〕 ソウル五輪

図 3.2－3　ＮＥＥＤＳ－ＩＲデータのキーワード抽出

する処理や，キーワード辞書にない重要語を拾い上げる手法の開発が必要であろう。

また，日本電信電話公社でも，キーワード辞書とのマッチングにより新聞記事からキーワードを抽出するシステムの開発を進めている（図3.2－4）。このシステムでは，記事中の用語の品詞を判定して名詞だけを抽出して，接辞処理や不要語削除を行い，重要語辞書やシソーラスとのマッチングを試みている。またキーワード候補語の出現頻度や位置情報（何番目の段落に出現するか）などを総合的に判定して重要度を判る研究開発も進めている。このシステムでは，キーワード選定の結果をディスプレイで表示してオンラインでポスト・エディティングが実行できるようになっているが，自動抽出によってある程度の品質のキーワードが確保できれば，ポスト・エディティングを急ぐ必要がない場合もあり，バッチによる処理も可能にしている。またシソーラスはディスプレイ上に表示されるとともに，ポスト・エディティングの間に，新しい用語を



(1) ／大手スーパー／，／○○ストアー／，・・・の名詞が抽出されます。
(2) ／大手スーパー／が付属語解析され，／大手／／スーパー／の2つに分けられます。
(3) ／大手／という単語は，ストップ・ワード解析により，不要語として取り除かれます。
(4) ／スーパー／という単語から上位語の／大型店／，／小売業／，／商業／がつぎつぎに抽出されます。
(5) ／○○ストアー／が固有名詞かどうか調べられます。

（／大手スーパーの○○ストアーは，・・・／の文章が記事情報として処理される様子）

*シソーラス辞書にない固有名詞とかトピック的な用語を収録してある辞書

（通研月報Vol. 36 №12 1983）

図3.2－4　電電公社のキーワード自動抽出の構成

追加するための随時更新ができるなど，マン・マシン・インターフェイスを重視して設計されている。

3.2.4 今後の課題および利用の方向

これまでに述べてきたことは，自動抽出の方法論として特に目新しいものはないかも知れない。しかし，繰り返すようだが必要なことは，現在の技術レベルの中でどれだけ自動化の精度を上げられるかという問題である。そのためには，現在の手法に改良と工夫とを積み重ねていくことが最も大切であり，そこからまた新しい方法論への道が開けてくるかも知れない。

この観点から，いま最も急ぐ必要があるのは国語辞書やシソーラスなど基本的な辞書の作成，あるいはそのためのデータの整備である。キーワード自動抽出システムの開発にあたって，まず先立つものは辞書であり，プログラムの作成よりも辞書づくりに多くの時間とコストを要する。総合利用システム全体の知識ベースとして整合性のとれた体系を持つことは望ましいことだが，内容分析など高度な処理のための知識ベースの設計にはまだ時間がかかると思われる。とりあえずキーワードの自動抽出システムとこれに必要な辞書の整備を行う人も一つの選択であろう。

キーワード自動抽出を行うことで，文章情報利用の面でいくつかの発展が期待できる。

まず情報検索では，文献単位の検索でなくパラグラフ単位の検索が可能になる。欧文では当然のことになっているが，フルテキスト入力とキーワード自動抽出により，利用者は膨大な資料や文献の中から，希望するパラグラフだけを入手できるわけで情報検索の効果が最も端的に表われる。法律の条文検索や判例検索，またある人物についての情報を収集する場合にも，自動抽出は欠かせない。

次に文章中の数値の利用が考えられる。文章情報中の数値には，体系

化された数値データベースでは取り扱えない。いわば方言隻句的な数値データで情報価値の高いものがある。これを利用するためには，数値とこれに係わりのあるキーワードを結びつけた数値情報を抽出することが必要だが，これにはキーワード自動抽出の技術が役立つ。ただし文章中から，いわゆる5W1Hの事実情報を抽出するには構文解析だけでなく意味解析も伴う必要があり，これはまだ先のことになろう。

分かち書きと漢字カナ変換を使って，文章の点字訳や音声への変換は，本来の開発目的から逸脱するが，こうした面への応用の道が開けるのも悪くないであろう。

3.3 文章情報内に出現する自然語の頻度分析

文章情報内に出現する自然語を対象として，定量的な解析を施し，主題内容を代表する候補語を抽出・分析する目的には，次の2つのレベルがある。

① 重要性の判別

② 順位性の判別

①の重要性の判別とは，その文章情報内でキー（Key）となる単語（Key-word）を判別することを目的とし，②の順位性の判別とは，①の結果を元に文章情報そのものの順位性（重要性）すなわち文章情報の配列を判別することを目的とする。

文章情報の素データとして利用するには，そこには目的を必要とする。目的のためには，他の因子が必要になる。目的を達成するには，目的とする因子データと上記分析結果との相関分析を必要とする。しかし，いずれの目的のためにも，素データとなる文章情報内の自然語の分析が必要となる。そこで，以下に文章情報内に出現する自然語の頻度数分析の実際的な方法について概説する。

なお，文章情報中に出現する自然語の頻度数を基とする，ここで主に述べる自動索引化以外の応用分野 — 自動抄録，自動分類，さらには，特定の主

題領域の専門用語の様態分析（専門用語の頻度数をベースとする専門用語のネットワーク化）による特定主題領域の事象遷移分析等への応用 — については詳しくは省略する。これらの応用分野については，対象とする文章情報の特性解析から検討をすることが必要になる。

3.3.1　頻度数分析によるキーワード抽出の目的

文章情報を対象に，コンピュータを用いて文章中に出現する自然語の頻度数分析を行い，自動的にキーワードを抽出する方法は，単に文章情報を処理する技法に留まるのではなく，図3.3－1に示すように，例えば文章情報の蓄積・検索システムの中心的な役割りを果たす機能として位置づけられる。その主たる目的は，大量の文章情報を対象に索引ファイルを自動的に作成するための手段といえる。そこでキーワードの頻度数分析手法は，この種の目的を達成するために，次の3つのレベルで研

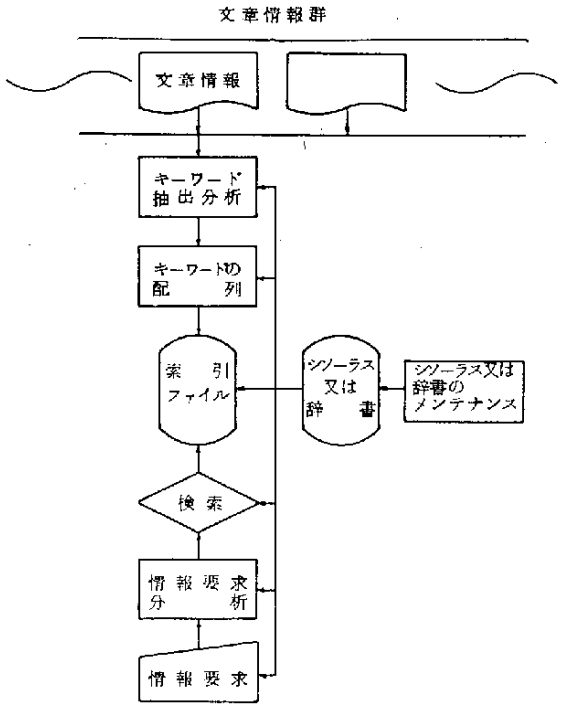


図3.3－1　情報検索におけるキーワード抽出処理機能

究・開発および実験が成されている。

① 文章情報の自動分析のための研究

② 索引ファイルの自動作成のための研究

要は，大量の文章情報を対象とするデータベースの自動生成のための研究である。

③ 文章情報の検索のための研究

①のレベルの研究は，文章中に出現する自然語の様態を分析する目的をもち，自然語の確率事象を分析する手法の研究である。

②のレベルの研究は，①の結果抽出されたキーワードの配列問題の研究であり，索引語数の最適容量化の課題を現在なお残している。

③のレベルの研究は，Salton，Gらの研究に代表されるように，情報要求との間に，要は情報要求の分析も含め，検索効率，すなわち検索精度の問題を含め，今なお研究が盛んに進行している。

しかし，文章情報を対象とするキーワードの頻度数分析手法の研究は，コンピュータ等分析ツールの発展により，上述するように単に文章情報の蓄積・検索システムの確立を目指すだけではなく，特に，①の研究成果を，例えば，文献（主に研究・開発および学術活動の成果である論文）を対象に応用することにより，研究動向の分析のために，さらに技術および製品開発の動向，さらには社会事象の動態分析，マーケッティング分析等に広範囲に応用されだしている。この種の目的を達成するために，最近では上述①のレベルでの分析結果を基に，個々の問題解析のための分析を目的的に利用する研究，例えば，クラスタリング分析，あるいは多変量分析および定量化分析等の手法の利用が活発化してきている。しかし，文章情報を対象に，この種の目的を達成するには，単にそこに出現する自然語の頻度数を対象に表層的に分析するのでは誤差が生じる。そこには，文献情報の主題内容を厳密に分析し，主題内容を表現する索引化が成されていなければならない。そこで以下に，まず文章情

報を対象として行う索引化の機能について概説し，次に文章情報を対象にキーワードを抽出するための基本を成す，上述①のレベルの頻度数分析の問題について考察する。

3.3.2 文章情報を対象とする索引化分析

一つの文章情報は，一般に以下のような構造から成る。

① 出典事項（ Imprint data ）……作成者（一般的には著者），作成年月日，掲載物名，ページ等

② 内容（ Content ）……情報内容を記述する文章

文章情報を対象とする索引化（ Indexing ）とは，上記構成要素から主題内容（ Subjects ）を抽出し，その結果を一般に索引語（ Index term ）によって表記することをいう。索引語によって主題内容を表記する理由は文章情報が起承転結を持つ文（ Sentence ）から成ることから，全文を読まなければ，その主題内容を一瞬にして把握しえないので，索引語で代表させることに基づく。

索引語は，一般に自然語形式で表記する。しかし，文章情報を，例えば“群”として処理したり，管理する場合の統一性を考慮し，表記形式の統一を図るために，例えばシソーラス（ Thesaurus ）とか辞書（ lexicon ）を利用し，その登録語（ entry term ）との照合を図り，表記の統一を図る。索引語は，さらに例えば分類番号に置換し利用する場合もある。

文章情報を対象とする索引化分析は，例えば，文章情報の蓄積・検索システムといったことを意図して行う場合には，文章情報の構成要素のレベルと同じく，次の２つのレベルで抽出し索引語を付与するのが一般的である。

① 出典事項を対象とする索引化……特に，文章情報と実際に活用する時に参考とする作成年月日，作成者の所属機関名等が重要な

被索引化項目になる。

② 文章内容を対象とする索引化……文章情報の内容を対象に主題事項を抽出するための索引化である。

ところで，文章情報から索引語を抽出するプロセスは，図3.3－2のようになる。

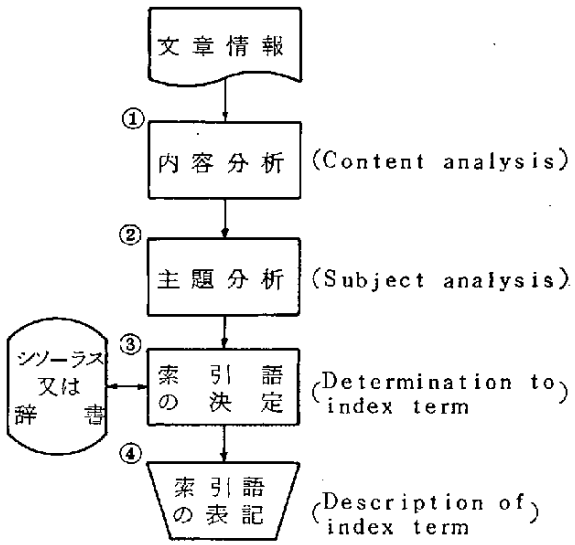


図3.3－2　文章情報を対象とする索引化分析のプロセス

① 内容分析（Content analysis）……文章情報を読み，理解し，内容を把握する機能。

② 主題分析（Subject analysis）……内容分析の結果に基づき，中心となる主題を抽出する機能。また，対象としている文章情報が有効に利用されると思われる領域（分野）の主題，さらには当索引語抽出機能を包含するシステムの目的から発せられる制限主題等を，現在対象としている文章情報との相関において判別し主題を決定するための機能。

③ 索引語の決定（ Determination to index term ）……抽出された主題概念を表わす単語を，シソーラスまたは辞書の登録語と照合し，索引語を決定する機能。

④ 索引語の表記（ Description of index term）……決定された索引語を，システムのデータ・フォーマットへ記述する機能。

文章情報を対象に，コンピュータを用いて自動的に索引語を抽出する方法として，１つの文章情報に対して，定性的に抽出する方式と，定量的に抽出する方式とがある。定性的な抽出は，文章情報を構成する自然語文の構造（文法）を解明し，自然語文の構成要素である自然語毎の機能を明らかにしていき，文章情報の内容分析，そして主題分析，さらに索引語の決定を行う方式である。定量的な抽出は，文章情報内に出現している自然語の統計的分析および確率的解析により，出現している自然語の重要度を評価し索引語候補語を抽出し，索引語として決定する方式である。ところが，主題索引化分析（ Subject indexing ）において，文章情報の内容表現が論理的な展開によって記述されていることから，主題内容を表現する索引語を，例えば，次のように構造的に抽出する必要がある。

① 実験対象動物名および属性値（性別，年令等）

② 実験方法

③ 使用薬品名

④ 使用器具

⑤ 反応色

⑥ 実験場所等

そこで文章情報を対象に前記のように構造的に索引語を抽出するには，自然言語処理（ Natural language processing ）および文章処理（Text processing ）を行い解析し抽出することが必要になる。このことに対して定量的な抽出方式では，出現する単語を統計的に測定し，確率的に

判定することに留まることから，索引語を抽出する根本的な手段にはならない。また，自然言語処理および文章処理による方式は，それ自体の技術の確立が完成していないことから利用できないでいる。そこで，現在，実際に稼動している文章情報の蓄積・検索システムを目的とする索引化は，人手による定性的な解析手段によって行われている。

しかし，文章情報を対象に出現している用語を定量的に解析する手段は，大量の文章情報の自動分類（ Automatic classification ）等に応用されているのは事実である。

3.3.3 頻度数分析によるキーワード抽出（一般論）

文章情報中に出現する自然語の頻度数分析によるキーワードの分析機能を図3.3－4に示す。キーワード抽出機能は大別して次の2つの系に分けられる。

① キーワード・ファイル（リスト）の作成機能

② 新規入力文章情報に対するキーワード抽出機能

①の機能は，目的とする主題領域の基本キーワード群を決定するために，その主題領域に関係する最適量の文章情報群[*1]を対象に予じめ分析するための系である。

②の機能は，①の分析結果によって作成されたキーワード・ファイルを基に，新規入力文章情報のキーワードを抽出するための分析機能の系である。

一つの文章情報中に出現する自然語を対象に，キーワードを抽出し，対象文章情報の特性を把握，または，対象文章中より“ある種の情報”

---

*1 最適量の文章情報群とは，分析対象となる文章情報の母集団の適性化のことを言う。一般に文章情報を対象とする集合，Bradfordの実験則に基づく，Core Text（核文章情報）として分析し用意する。

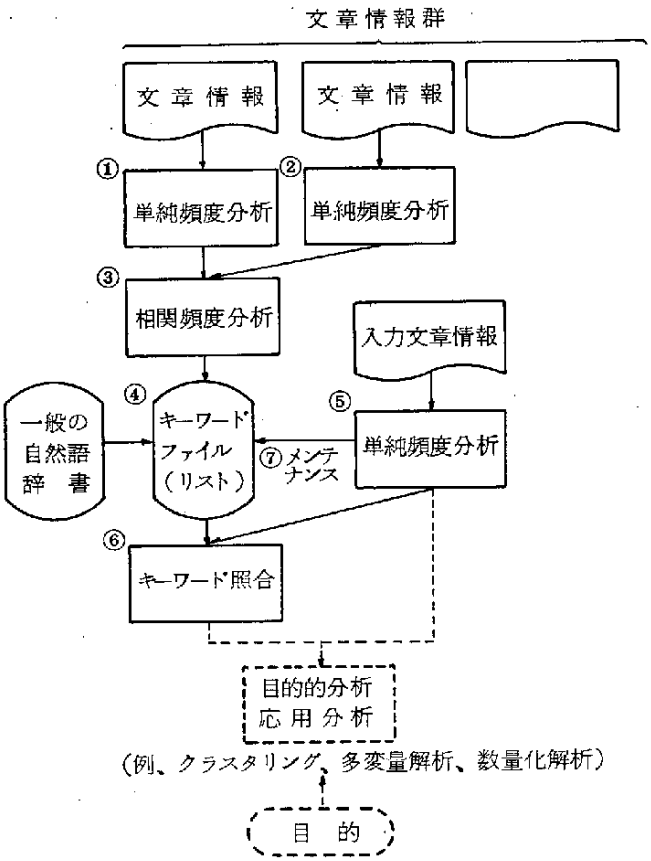


図 3.3－3　文章情報の頻度数分析によるキーワードの抽出機能

を抽出するには，その分野で中心を成す，あるいは，その分野の特性を代表するキーワードを事前に決定しておき，新規入力文章情報中に出現するその種のキーワードの頻度数を統計的に分析し抽出する必要がある。一つの文章情報中に出現する自然語の頻度数分析の結果だけでは，その分野の中で相対的に文章情報の特性を把握したり，あるいは“ある種の情報”を抽出することはできない。そこで，まず前述①のキーワード・ファイル作成のための機能系が必要になる。以下，分析のプロセスについて概説する。

① ある目的のために収集した文章情報群の中からキーワード候補語を抽出するために自然語の頻度数分析を行う。文章中に出現する自然語の頻度数分布は図3.3－4に示すような指数分布になる。

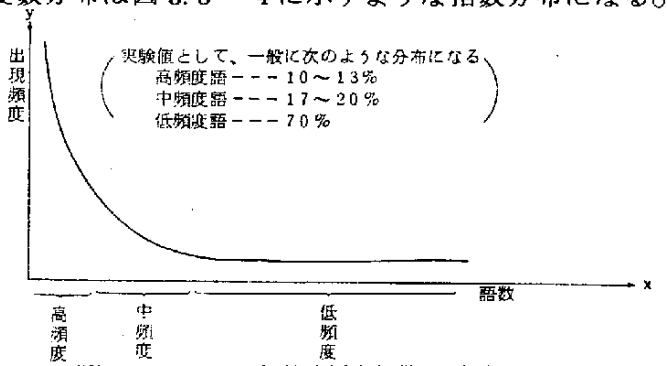


図3.3－4　自然語頻度数の指数分布

図3.3－4の分布の自然語を，Luhn，H.P.が英語について観察した結果，次の3つのグループに分けられることを提示した。[1)]

㋑ 高頻度の所に分布する語は，英語文中の機能語（例　a，an，of，for等の約1,000語）が多く，文章特性を特に示す語とはならず，これらの語を文章情報を解析する上での不要語（Stop word）と見なしうる。

㋺ 次のレベルの高頻度の所（中頻度）に分布する語は，㋑の不要語とは異なり，英語文中の有意味なる語ではあるが，広く一般に使用される語であって，特に文章特性を示す語とはならず，これらの語を共通語（Common word）と見なしうる。

㋩ 低頻度の所に分布する語は，専門性の高い有意味語であることから，これらは重要語（Key word）と見なしうる。

そこで，この考え方を基に，まず，一つずつの文章情報内の㋩でいう所のキーワードの相対頻度数を算出する。（①，②の機能）

㋩でいう所のキーワードはLuhn，H.Pの分析の結果，前述のキーワード・ファイル内に登録されるキーワード候補語と見なすことができる。

② キーワード決定のための相関分析（③の機能）

そこで，もう少し統計的に有意味なレベルのキーワードを抽出するには，目的とする主題領域に関連する最適量の文章情報個々との相関性を分析し決定づける必要がある。

文章情報内のあるキーワード候補語の相対頻度数をｆ（０＜ｆ＜１）とし，そのキーワード候補語が，他の文章情報内で使用されているときの相対頻度数をｒ（０＜ｒ＜１）とすると，ｆとｒとの相関値の程度，すなわち，キーワードの相関係数（ＡＦ：Association Factor）によって，両方に関連するキーワードを選定する。

相関係数は，下記３通りの方式で算出が可能である。

㋑ $AF=\frac{f}{r}$ ……比率により

㋺ ＡＦ＝ｆ－ｒ ……頻度表により

㋩ ＡＦ＝ｄ（ｆ，ｒ）……関数形で

要は，どの算出方式にしろキーワードとして決定づけるには，相関係数ＡＦの値の判定規準が問題になる。

③ キーワード・ファイルの作成機能（④の機能）

相関係数ＡＦの判定の結果により決定されたキーワードを，次の表記上の統一を図るために一般の自然語辞書とつき合せキーワード・ファイルとして登録する。[*2]

---

*2 英語文の場合，ここでさらにLuhn,H.P.が提示するところの「resolving power」の考え方による解析が必要になる。すなわち，上述迄の方式は，１単語を対象とする解析であって，複合語レベルの解析はなされていないことになる。キーワードは，１単語とは限らない。例えば，"Computer system" のように複合語が対象になる。そこで「resolving power」のレベルでの解析が必要になる。「resolving power」とは，例えば"Analysis of information"のように前置詞によって複合化する語を一つのキーワードと見直すとき，自然語と自然語との間に挿入される語を何語迄許しうるかという値のことを言う。
この事に対して，日本語文の場合には，"の"格が相当する。しかし，日本語文を対象とする場合，解析前の処理として原データが分かち切りされていない場合には，原データに対する単語切り，すなわち形態素解析（Morphological analysis）が問題になる。

④　新規入力文章情報に対する分析

以上の結果，用意されたキーワード・ファイルを基に，新規入力文章情報の解析が可能となる。[*3]

ところで，文章情報中に出現する自然語の頻度数を基に，目的的に例えばクラスタリング分析，多変量解析，さらには数量化分析を行うには，単に自然語の頻度数だけを対象に行うのでは不可能である。他の目的とする因子（例えば，文章情報の発行年，発行地，作成者の所属する機関名等）との相関において分析を並行して行うことが必要である。

3.3.4　キーワード抽出のための相関分析

目的とする主題領域に関連する文章情報群内のある一つの文章情報中に出現するキーワード候補語の相対頻度数を基準に，異なる他の文章情報のキーワード候補語の相対頻度数との相関係数を算出し，キーワードを決定していく一つの方式「相関分析による方式」について概説する。

A…aというキーワード候補語を持つ文章情報の数（但し，A＞0）

B…bというキーワード候補語を持つ文章情報の数（但し，B＞0）

f…a，b両方のキーワード候補語を持つ文章情報の数

（但し，f＞0）

＊3 新規入力文章情報の自然語の単純頻度分析の結果（⑤の機能）を，キーワード・ファイルと照合することによって，この入力文章情報のキーワードを決定する（⑥の機能）。また，新規入力文章情報から算出されたキーワード候補語が，キーワード・ファイル内のキーワードと照合しないことが，例えば，文章情報が時系列的に発生する場合（例，文献情報）には充分起る。この場合，例えば，キーワード候補語の人間系での評価の結果を基にキーワード・ファイルに新規キーワードとして登録する必要がある（⑦の機能）。

N…総文章情報の数

とすると，キーワード候補語aの相対頻度数Aが，キーワード候補語bの相対頻度Bに対して無関係に起きる確率P（A）は一定であり，またAに対して無関係に起きる確立P（B）も一定である。そこでA，Bの共出現の確率は，次式のようになる。

$$P(AB)=P(A)\cdot P(B)\cdots\cdots①$$

ここで，P（A），P（B）の文章情報の集合Nの中での期待値は次式のようになる。

$$Pe(A)=\frac{A}{N}$$

$$Pe(B)=\frac{B}{N}$$

故に，A，Bの共出現の期待値は，①式と同様に次式のようになる。

$$Pe(AB)=Pe(A)\cdot Pe(B)$$

$$=\frac{A}{N}\cdot\frac{B}{N}=\frac{A\cdot B}{N^2}$$

そこで，a，bの共出現頻度数fの期待値Fabは次式のようになる。

$$Fab=N\cdot Pe(AB)$$

$$=N\cdot\frac{A\cdot B}{N^2}$$

$$=\frac{A\cdot B}{N}\quad\cdots\cdots\cdots\cdots②$$

ここで，文章情報の集合Nの中に出現するキーワード候補語a，bの相対頻度A，Bが，独立変数として互いに無相関であることから，X^2（カイ）検定を行う。但し，a，b 2つのキーワード候補語の相関を検定するために，属性総計表として，それぞれの相対頻度A，Bを基に，次のように2×2の分割表に表わし計算する。

A ＼ B	b（B）	$\bar{b}$	計
a（A）	AB	B−AB	B
$\bar{a}$	A−AB	N−A−B+AB	N−B
計	A	N−A	N

故に，

$$X^2=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(f-Fab)^2}{Fab}$$

$$=\frac{(\frac{AB}{N}-AB)^2}{\frac{AB}{N}}+\frac{(\frac{B(N-A)}{N}-(B-AB))^2}{\frac{B(N-A)}{N}}$$

$$+\frac{(\frac{A(N-B)}{N}-(A-AB))^2}{\frac{A(N-B)}{N}}$$

$$+\frac{(\frac{(N-A)(N-B)}{N}-(N-A-B+AB))^2}{\frac{(N-A)(N-B)}{N}}$$

$$=\frac{N(AB(N-1))^2}{AB(N-A)(N-B)}$$

ここで，f＜5の場合にはYatesの補正を行う。

故に，$$X^2=\frac{N(|fN-AB|-\frac{N}{2})^2}{AB(N-A)(N-B)} \quad \cdots\cdots③$$

となる。

Stiles, H.E. はキーワード候補語a，bの相関係数FAとして③式の常用対数を用い，約100,000件の文献情報を対象に分析を行っている[2]。その結果，文献情報中に存在するキーワード候補語のa，bの関係について，次の2点を指摘している。

① AF＞fNならば，a，bの語関係は浅い。

② AF＞1ならば，a，bは無関係である。

以上のことから，相関分析においては1＜AF＜fNの相関性において，文章情報のキーワード候補語をキーワードとして，順次判定していくことが可能であると言える。

3.3.5 文章情報集合に対する頻度数分析

これまで述べてきた新規入力文章情報を対象とするキーワード抽出分

析法に対して，以下で，例えば，文章情報の集合（データファイル）に対して直接検索するシステムにおいて，文章情報の集合を対象にキーワードを抽出するための算出法についてSalton,G.[3]のレビューを基に紹介する。

以下では，文章情報 i に出現する自然語 k の頻度数を f_i^k として，自然語 k のN個の文章情報の集合合計頻度 F^k をベースに述べる。

この種の算出法については，既に各主題領域で導入されたキーワードリストの評価において議論されており，算出法それ自体の議論は，上述変数の範囲内では大差ないものとなっている。従って，ここでは以下2点について，例示的紹介にとどめることにする。

① 通信理論に基づく算出

信号Sを文章情報の集合中のキーワード候補の集合，ノイズ N^k を非キーワード候補語の集合とみなすと，以下のように通信理論の算出法が応用できる。

n個の文章情報の集合に対して，自然語 k のノイズ N^k は，次のように定義できる。

$$N^k = \sum_{i=1}^{k} \frac{f_i^k}{F^k} \log \frac{F^k}{f_i^k}$$

したがって，信号 S^k は，次のようになる。

$$S = \log F^k - N^k$$

文章情報の集合 k の中のあらゆる文章に，ある自然語が同一回出現しているとすれば，

$$f_i^k = 1$$

となり，

$$N^k = \sum_{i=1}^{n} \frac{1}{n} \log \frac{n}{1}$$
$$= \log n$$

となり，ノイズが最大となり，信号はゼロになる。逆に，ある自然語

の頻度数F^kで，ある特定の文章情報だけにしか出現しなければ，

$$N^k = \frac{F^k}{F^k} \log \frac{F^k}{F^k}$$
$$= \log 1 = 0$$

となり，ノイズがゼロとなり信号は最大になる。故に，ノイズおよび信号とキーワード候補語との間には関連性が存在することになる。

② 平方偏差に基づく算出

$\bar{f}^k$ をn個の文章情報における自然語kの平均頻度数とすれば，平方偏差$(V^k)^2$は，次のようになる。

$$(V^k)^2 = \frac{\sum_{i=1}^{n} (f_i^k - \bar{f}^k)^2}{n-1}$$

この平方偏差による算出法は，集合内に自然語kが均一に分布している場合，つまり，全てのf_i^kが同じ大きさの場合は，$(V^k)^2$が小さくなり，対応する用語はたいして有用でないことを示している。一方，kが稀な語であって，少数の文章情報にしか発生していない場合には，大部分のf_i^kがゼロとなり，$(V^k)^2$は小さくなり，この方法は，キーワードを特定化するに必ずしもよい算出法にはならない。すなわち，この算出法は，f_i^kが平均的である場合に有効となる。

3.4 米国における文章情報処理先進事例調査

文章情報データベースの総合的な利用のためには，国内のみならず，海外情報の有効活用を図る必要がある。これら情報の総合解析システムの研究開発にあたって，言語をより機械的な処理により翻訳して利用する方法を考えることは，さけて通れない関門であると言える。

歴史的にいくつかの言語を用いている国や国際機関等においては，異言語による障害により迅速に，高度に対応する必要上からも，コンピュータによる翻訳等の研究が，早期より進められてきた。これら先進的な研究事例を調

査し，本プロジェクトのめざすシステム開発の参考とするべく，本年度は米国を中心に，自然言語処理の研究としてハーバード大学，イエール大学，モントリオール大学の調査，また，文章情報データベースの現状調査としてＩＳＩ社，ＵＣＬＡの調査結果を紹介する。

3.4.1　自然言語処理の研究

《その１》　ハーバード大学における自然言語処理研究

１９６０年代初めにSusumu Kuno, Anthony G.Ottinger らにより作成された自然言語処理のためのシンタクティック・アナライザは，先駆的な構文解析システムであるというにとどまらず，非常に広範囲の英文に対応する事を目的とした文法規則，統語論的な構文構造を把握する為の各種補強メカニズムを備えた総合的システムであるということは比較的知られていない。当時のコンピュータ能力などの問題から，この研究は中断状態であったが，最近になって再び見直されてきており，久野氏自身も文法の拡充などに向いはじめている。

(1)　パージング・メカニズム

予測型文法規則（Greibach 標準形の文脈自由文法規則）を用いて，文の左から右へバックトラックしながら進む典型的なトップダウン・パーサである。その後，文全体あるいは部分構造のアンビギュイティに起因する解析所要時間の爆発を抑えるため，くり返しパスの除去のメカニズムが加えられた。[*1] これは，一度成功した部分構造（WFS：Well-Formed Substring）を記憶し，再試行に際してはその中に立ち入らないというものを中心とした手法であり，ボトム・アップ手法で用いられる中間ノードの記憶方式に対応するものであるが，

---

*1 Kuno, Susumu : The Predictive Analyzer and a path Elimination Technique, Communication of ACM (1965)

完全なリダンダンシーの除去は不可能である。他方，記憶容量の超過はありえないものとなっている。また，数の一致のテストなどの補強メカニズムはこの処理と矛盾するので，後の処理に廻され，パージング自体は文脈自由文法規則の枠だけによるものになっている。

(2) 文法規則

前述の様に，予測型文法規則の形式で書かれている。種々の分野の英文に対応する為に規則は非常に細かいものとなっており，規則数は約2,５００程度になっている。また，パーサの終端記号を表わすホモグラフ・コード（ＳＷＣ：Syntactic Word Crass）は約１８０種，非終端記号（プレディクション）は約９０種ある。各々規則の拡充とともに，整理・追加が行われている。

(3) 文法記述における略記法

パーサは，２種類の略記法に対応することができる。

一つには，andにより同じ種類のプレディクションがもう一つ発生しうることを同一規則で記述することをdroppable prediction という特別なプレディクションを用いることで可能にしている。

次に，リスティング構造（“A，B，……〔，〕and C”の並列構造）に対応するためのダミー・プレディクションがある。これは，リスティング構造が，任意個の同一レベル・プレディクションを並列結合しうることに対応したメカニズムであるが，文脈自由型文法の枠内からはとび出すインプリメントとなってしまうようである。

(4) 数の一致テスト

条件節の中を例外として，一般に述語主語と述語動詞との間には数の一致が必要である。このパーサでは，入力記号や規則のもつ数の情報を後のプレディクションに転与する種々のメカニズムにより実現している。ここでは，トップダウン手法のため多少こみいった処理が必要となっている。

(5) シンタクティック・ロール・ワード

各語に適用された規則の統語的属性を割り当てるメカニズムである。この性質は，その規則が適用された環境によっても違うことがあるので，プレディクションの細分類コードも用いられる。この細分類は親ノードから情報を受け継ぐ形で実現される。これにより，同じ主語の規則であっても，主節の述語に対応する主語である云々という詳細な区分ができる。

(6) センテンス・ストラクチャ・コード

規則が標準形で書かれているために，パージングから自然に生まれる木構造は，本来的な構造から見るとくずれた形となってしまう。現在，最左導出型規則で抜け落ちているプレディクションを補充するようなメカニズムと規則の補充も試みられているが，ここでは，元々のシステムで実現されていたセンテンス・ストラクチャ・コードのメカニズムについて説明する。

例えば，次の文の分析があったとする。

〔〔 They 〕$_{主}$〔 are〔 flying planes 〕$_{補}$〕$_{述}$・〕$_{文}$

これは，センテンス・ストラクチャ・コードで，次の様に表現される。

They	1 S
are	1 V
flying	1 N A
planes	1 N
・	1 .

これにより，“ flying ”は平叙文（“ 1 ”）の名詞補語（“N” ― 主名詞は“ planes ”）を修飾する限定詞（ “ A ” ）の役割をもつなどのことが表現される。

各語のもつ部分構造（“ flying ”であれば，“ N A ”は必要ならプ

レディクション細分類コードも合わせ用いて，適用された規則により決定される。その部分構造が挿入されるレベルは，そのプレディクションを生成した規則が与える相対レベル値（シフティング・コード）によって決まる。たとえば，上の例で動詞句プレディクションに対する“ are ”に適用された規則：＜動詞句プレディクション＞→＜be＞・＜名詞補語プレディクション＞に対しては，この語位置にストラクチャ・コード“ V ”を与え，生成するプレディクションには同一レベルという意味で，シフティング・コード0を与えている。

フローティング・ストラクチャについては，修飾先が決定できない場合がある。たとえば上の例で，“ planes ”と“・”の間に前置詞句が入ると，それが形容詞的に“ N ”の下につくか，副詞的に“ 1 ”の下につくか不定である。このシステムでは，直前の修飾可能な最下位レベルにかけているとのことである。

(7) フローティング・ストラクチャ

前置詞句や副詞（句）のように，その環境とは比較的独立に，文中のいろいろな位置に出現することができる構文単位をフローティング・ストラクチャと呼ぶ。特に，それが周囲のどれを修飾しているかを文法規則が決定していないという意味でも“ floating ”な構造である。この文法では，フローティング・ストラクチャに関する規則が相当の部分を占めている。

(8) 判定されないアンビギュイティ

規則が予測型に書かれ，左から右に順次決定する形となっているため，修飾に関するアンビギュイティの相当部分はペンディングとなっている。たとえば

（形容詞）・（名詞）・（名詞） and（名詞）（形容詞節）

において，可能な多くの解釈に対して，構文解析では1つの解しか与えない。

このように，ペンディングにしているアンビギュイティがあるために，パージングの負担は軽くなるが，後でこのアンビギュイティをパーサが別の解としたアンビギュイティとともに解決する必要があるときに問題が起こるかもしれない。

《その２》イエール大学における知識情報処理研究

(1) 調査の背景

自然言語処理の方法論は，チョムスキーに始まる句構造文法に基づく構文解析法から，動詞の各支配に基づく処理，ＡＴＮ（Augmented Transition Network），拡張ＬＩＮＧＯＬ，モンターギュ文法に基づく処理，あるいは推論をも可能とするような意味表現形式の研究というように，言語が表わしている深層の意味レベルにまで接近しつつあるような感さえある。

こうした方法論の推移は，自然言語を特定のメカニズム（たとえば構文解析法）でとらえようとするとき，それからはみ出す事象が多々あり，またそれを解析するためには深層の意味レベルまでの理解を要請するような事象がしばしば存在するといったことにも起因している。

いずれにしても，ある特定のメカニズムを採用したとき，実際の言語現象に対してある部分をすくい上げてある部分を捨象するという事になる。すくい上げる部分のメリットが多ければ，その手法は実用システムのアルゴリズムとして利用される。

そうした実用システムへの応用という面では，ようやく構文解析法がその対象になったというところである。

しかし一方では，構文解析法の功罪は，既にいろんな所で研究・実験されたように，厳然として存在している。実用システムへの適用という面でその功罪をあげてみれば，まず主なメリットは，

① アルゴリズムが明晰である。

② 文法（構文規則）が理解しやすい。

③ 辞書データが作成しやすい（基本的には品詞情報を付与すればよい）。

等々である。一方，デメリットは，

① 多数の解（ambiguity）が生じてしまう。

② 理解レベルが表層に近い。従って，推論等の対象とするには，解析が荒すぎる。

等々である。

これに対して，イエール大では，言語の表層というより深層の意味理解を目指す方法論をとっており，構文解析法とは好対照であるといえる。そこで，ここでは実験システムへの応用という点を念頭におきつつ，彼らの方法論及び現状を調査した。

(2) 調査内容

まず，自然言語処理に対する彼らの研究の位置づけに対しては，次の通りである。

① 一時的な単純なシステムを構築すると考えてはいない。一般的なことを理解する一部として自然言語理解を考えている。

② 新聞等をマシンで処理するということを考えたとき，一般にはキーワードを記事に付与してデータベースを構築し，キーワードでデータベースを検索するという方法をとっているが，実はこれは非常に誤解が多い。たとえば，「フォード」というのは「フォード大統領」か「フォード自動車」かあいまいである。そこで，こういう立場はとらない。

③ 言語の背後にある意味を理解することを目指す。そのため言葉自体は無視する。これは「世界」を理解することとも言える。いかにしてマシンが知識を吸収していくかに興味をもっており，そのためには言葉にとらわれないレベルでの理解が重要だ。

④ 従って，一般の言語表現は「意味」を表現しているというたてまえにたって，文や語を解析するという立場をとる。

⑤ sentence というより story を理解していくためにも，意味の理解は不可欠である。

ひと言でいえば，言葉の表層より深層の「意味」理解に重点をおくということであり，研究のスパンも５～１０年の規模で考えている。

こうした考え方から，現在の主流と考えられる翻訳プロジェクト（ＥＣやモントリオール大やグルノーブル大等）のアプローチに対する評価も，「syntax－basedである。すなわち言語の形態に関心をもって意味には注目していない。そのため，他の言語ペアに対しては，また別のプログラムで対応することになる。これに対して，イエール大では意味の把握を行うため，そのレベルからはどの言語も生成できる」となる。実際にイエール大では，原文がスペイン語の記事を意味レベルでとらえ，それから英語とドイツ語に翻訳している。しかし，これは一方は翻訳のための実用システムを目指しており，他方はもっと長期のスパンでの言語理解を目指しているため，単純な比較はできない。

イエール大における意味理解の方法論の中心は，「概念依存理論（Conceptual Dependency Theory）」であるが，その方法論の内容及び考え方に対しては，次の通りである。

① Concept から意味を理解する。

② 構文解析は前段階として実行したりせず，意味理解のプロセスで同時に実行する。言葉の順序は重要であるが，それ以上にConcept そして文の意味が重要である。

③ Syntax は，言わばComputer language でありComputerにとって処理しやすいとは言えるが，Natural language ではない。

④ 人間の記憶の構造，アクセス，再組織化，学習過程といったものが大事であり，それがあって新説への適応もできる。

⑤　意味理解の中心としているものは，１２種類の基本動詞（後述）である。

⑥　動詞だけでなく名詞の taxonomy も大事であるが，あまり行われていない。

⑦　concept には，小さい概念から大きい概念まで様々ある。

⑧　推論まで考えたとき，conceptual dependency レベルで考えておけば，言語を問わず処理できる。

⑨　聞き手によって違った表現を使うことに関する研究も行っている（例えば，聞き手が警察官のとき「犯人」と言い，テロリストのときは「ヒーロー」と言う）。

イエール大における実験システムに用いている記事は，ＵＰＩ通信からのテロリズムに関連する記事であり，その数はあまり多くないようである。

また翻訳システムに関しては，実際はデモをみせていただいたが，システムの規模は次の通りである。

①　対象記事はテロリズムに関連する記事（スペイン語）で総数５０記事（１記事あたり１～２文）。

②　システムは意味表現された中間表現（概念依存）から英語に翻訳するもの及びドイツ語に翻訳するものであり，開発言語はＬＩＳＰである。

③　concept はＬＩＳＰ言語で８５１行で記述されている。生成規則は concept ごとにある。

その他，ＵＰＩ通信からのニュース記事の内容を要約して記憶していくＦＲＵＭＰというシステムがあるということである(5)入手資料を参照）。

(3)　課　　題

言語の意味表現ということに関して，イエール大のアプローチは示

唆的ではあるが，実用システムへの応用という面からみると，まだ見えにくい面が多い。具体的に言えば，概念依存理論を使って，現実の言語現象にたえうるだけのデータ（ conceptを記述する規則等）をそろえきれるかという点である。たとえば，概念依存の骨子となっている「概念カテゴリ」の集合が６種，「概念依存関係のシンタックス」が１６種，「基本動詞（ Primitive　ＡＣＴ）」が１２種あり，これらが様々な組合せで登場する。これを，更に細かくみれば次のような問題がある。

① 動詞をいくつかのPrimitive　ＡＣＴ，対象及び状態情報を使って，個別に記述しなければならない。

② 概念カテゴリの基本である「ＡＣＴ」に対してPrimitive ACTを考えたように，他の概念カテゴリ「ＰＰ」（現実世界の対象），「ＰＡ」（対象の属性），「ＡＡ」（行為の属性）等に対しても，そのPrimitive　を設定していく必要があろう。

③ このように概念依存による表現を現実世界に適用していけば，Primitive　ＡＣＴ，「ＰＰ」，「ＰＡ」，「ＡＡ」等の組合せが相当に多くなる恐れがある。特定分野に限ったとしても，果して記述しきれるか。

以上は，実用システムへの応用という面からみたときの否定的側面であるが，一方では，構文情報を使った解析によるambiguityの爆発等に対する解決へのヒントや，文章情報の内容理解及びその表現方法に対するヒントも含んでいる点も見逃せない。

(4) 概念依存理論

以下に，イエール大の自然言語処理の中心理論である概念依存理論（ Conceptual Dependency Theory ）について簡単に述べる。（詳細は，文献「Conceptual Information Processing」参照）。

まず，その着想は，言語に依存しない意味表現を求めるため，文を

下位概念の構成要素に分解するということであり，その記述を「概念カテゴリ」の集合と「概念関係のシンタックス」によって行う。

概念カテゴリは，次の６種である。

① ＰＰ　：現実世界の対象（例　「John」，「book」）

② ＡＣＴ：現実世界の行為（例　Primitive ACT参照）

③ ＰＡ　：対象の属性　（例　「health」）

④ ＡＡ　：行為の属性

⑤ Ｔ　　：時

⑥ ＬＯＣ：場所

概念関係のシンタックスは１６種あるが，そのうち代表的９種を次に示す。

① ＰＰ⇔ＡＣＴ：ある行為者（ＰＰ）が，ある行為（ＡＣＴ）を行うことを示す。

② ＰＰ⇔ＰＡ　：ある対象（ＰＰ）がある属性（ＰＡ）をもっていることを示す。

③ ＡＣＴ$\overset{O}{\leftarrow}$ＰＰ：行為（ＡＣＴ）の対象（Object　）がＰＰであることを示す。

④ ACT$\overset{R}{\leftarrow}$PP$_2$, ACT$\leftarrow$PP$_1$：行為（ＡＣＴ）における対象の主体（PP_1）と客体（PP_2）を示す。（ＲはRecipientの略）

⑤ ACT$\overset{D}{\leftarrow}$PP$_2$, ACT$\leftarrow$PP$_1$：行為（ＡＣＴ）における対象の方向（Direction）がPP_1からPP_2であることを示す。

⑥ ＡＣＴ$\overset{I}{\leftarrow}$⇕　：ある行為（ＡＣＴ）に対する助格（手段や方法）を示す。（ＩはInstrumentalの略）

⑦ $\overset{X}{Y}$⇑　：概念Ｘが概念Ｙをひきおこしたことを示す。

⑧ PP$\Leftarrow$PA$_2$, PP$\Leftarrow$PA$_1$：対象（ＰＰ）の状態がPA_1からPA_2に変化したことを示す。

⑨ $PP_1 \leftarrow PP_2$ ：PP_2はPP_1に含まれることを示す。

文をこれらの概念依存で記述するためには，更に各Wordsをこれらの概念の組合せで記述する必要があるが，その基本となるものはＡＣＴであり，そのＡＣＴは次の１２種類のPrimitive　ＡＣＴに分類される。

①ATRANS：所有や制御のような，抽象的関係の変換。

②PTRANS：対象の物理的位置の変換。

③PROPEL：対象に対する物理的力の適用。

④MOVE　：動物の身体の一部の移動。

⑤GRASP　：行為者による対象の把握。

⑥INGEST：動物による対象の収容。

⑦EXPEL　：動物の身体から外界への放逐。

⑧MTRANS：動物間または動物の内部における精神情報の変換。

⑨CONC　：動物によるアイデアの概念化あるいは思案。

⑩MBUILD：動物による古い情報から新しい情報の構築。

⑪ATTEND：感覚器官を対象に向ける行為。

⑫SPEAK　：口から音を発する行為。

これらを使って，個々の語（Word）が記述されるが，その中心となるものは動詞であり，動詞はある特定の関係における１つ以上のPrimitive　ＡＣＴと対象と状態情報を使って表現される。

また，概念依存で使用されるその他の規則として，次のものがある。

① Objective, Recipient, Pirective そしてInstrumentalという４種の概念的格がある。

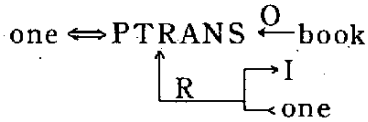
② 各ＡＣＴは，これらの格のうち２～３個を必ずとり，かつ省略できない。

③ Instrumental 格は，それ自体ＡＣＴとその格を含む概念化である。

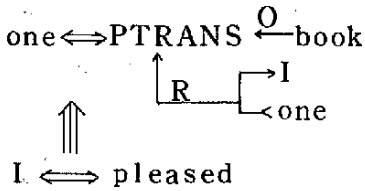
④ 生物対象のみが行為として作用する（ただし，PROPELは除く）。

次に，簡単な例として2，3あげる。

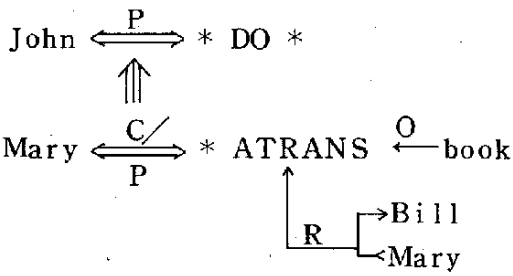
① I have a book



② I want to get a book



③ John prevented Mary from giving a book to Bill.



(注) ⟺の上の「P」は過去，「C」は仮定，「／」は否定を表わす。

(5) 入手資料

入手資料とその概要を以下に示す。

① 「Skimming Stories in Realtime :
An Experiment in Integrated Understanding」
Gerald Francis Dejong Ⅱ
May 79 Research Report #158

＜概　要＞

自動的なテキスト分析の新しい方法論を述べる。ＵＰＩ通信からのニュース記事を skimming して，要約を蓄積する。ニュース記事を正確に処理することができる。システムはFRUMP（ Fost Reading Understanding and Memory Program ）という名称でインプリメントされている。

② 「Conceptual Intormation Retrieval」

Roger C. Schank, Janet L. kolodner, Gerald Dejong　　Dec.80 Research Report ＃190

＜概　要＞

知的検索システムにとっては，自然言語を理解し，記事を構築あるいは再構築し，その記憶をサーチする知的なヒューリスティックを使う能力が必要である。こうしたシステムは，テキストと自然言語による質問の両方を理解する必要がある。これを合理的に行うためには，テキストの概念的内容とテキスト理解に必要な知識を適切に構成しなければならない。ＣＹＲＵＳとＦＲＵＭＰシステムは，こうした能力をもっている。ＦＲＵＭＰはＵＰＩ通信からのニュース記事の概念的内容を分析して記事の概要（重要人物についての記事の概要）を作り，それをＣＹＲＵＳに送る。ＣＹＲＵＳは，自動的にその記事の内容を蓄積して，自然言語によってなされた質問文に対して情報を検索できる。

③ 「Memory, Meaning, and Syntax」

Roger C. Schank, Lawrence Bimbaum

Nov.80　Research Report ＃189

＜概　要＞

Syntax 及び意味の役割について概説する。意味と世界に関する知識は，言語理解の第一段階においても重要な役割を演ずる。これ

は，統語上の知識が言語を処理する際に，特別の役割を果たさないことを意味する。

④ 「Representing Meaning : An Artificial Intelligence Perspective」
Roger.C.Schank
Apr.81 Cognitive Science Technical Report #11

<概　要>

意味に関する一般的な論文。

⑤ 「Modeling Memory for Language Understanding」
Roger C.Schank, Mark Burstein
Feb.82 Research Report #220

<概　要>

コンピュータによる自然言語理解の研究によって，理解のプロセスにおいて記憶の本質と構成が中心的な役割を果たしていることがわかった。本書では，言語と記憶プロセスが徐々に増大していくIntegrated 機能の一連のコンピュータ・モデルを概説する。

⑥ 「Representation and Translation」
Steven L. Lytinnen, Roger C. Schank
May.82 Research Report #234

<概　要>

翻訳システムにとっての知識の重要性。

《その3》モントリオール大学の機械翻訳システム

モントリオール大学においては，１９６５年に設立されたＴＡＵＭグループ（ Traduction Automatique de l'Université de Montréal ）により機械翻訳システムの研究開発が行われていた。ＴＡＵＭグループは，一貫して高いレベルでの研究を続け，実用システムとして天気予報に関

する英仏翻訳システムＴＡＵＭ－ＭＥＴＥＯ（１９７５～７６開発）を実現し，その後，航空機の保守マニュアルをサンプルとして用いたＴＡＵＭ－ＡＶＩＡＴＩＯＮの研究開発を行っていた。しかし，ＴＡＵＭ－ＡＶＩＡＴＩＯＮが一定の成果をあげたところで，カナダ政府はプロジェクト予算をストップし，研究は中断してしまっている。ＴＡＵＭのメンバーも，引退し，あるいはグルノーブルなどに参加するなど分解状態である。政府としては，他の市販翻訳パッケージなどの採用を考えていたようであるが，再び数年のうちに，何らの形で再組織したいとの意向でもある。しかし，長年にわたって培ってきたグループも，一旦分散してしまうと，再び結集するのは非常に難しいとのことであった。

(1) ＴＡＵＭの基本思想

モントリオールは，英語圏のカナダにあって，フランス語を公用語とするケベック洲の首都であり，英仏翻訳システムに関する高いニーズをもつ。他方，それだけに翻訳の質においても高いレベルが要求されるようで，ＴＡＵＭのメンバーも，既存の翻訳システムに対しては，WeidnerやＡＬＰＳなどは翻訳システムというよりはワード・プロセッサ程度のもの，ＳＹＳＴＲＡＮなどは第一世代の翻訳システムといった調子で，余り高い評価は与えていない。このように，ＴＡＵＭの考える実用レベルとは，相当に翻訳精度の高いものであるが，これを実現するために，まず対象の言語範囲を限定する。ＴＡＵＭ－ＭＥＴＥＯが成功したのは，天気予報という，そこで使われる語の範囲も文型も非常に限られたものであるという特殊性があったと云える。ＴＡＵＭ－ＡＶＩＡＴＩＯＮにおいても，航空機マニュアルの水力学に関する章だけを対象としてシステムの構築を図ったということで，こういった言語範囲の限定のことを，彼らはサブランゲージという表現を用いている。このように限定したサブランゲージに対して，統語論的なものは勿論として，意味論的な処理も導入することが必須と考

えている。

ＴＡＵＭのシステムとしての特徴は，典型的なトランスファ方式を用いていることにあろう。これは，現在各国で行われている機械翻訳の研究開発の多くにおいて採用されている方式で，その意味でもＴＡＵＭの果した功績は大きい。システム的には，また，各フェーズを直接プログラミング言語で記述するのでなく，各々の目的向けに，専用の形式言語を開発して用いているという特色があり，翻訳や言語学の専門家がコンピュータ言語に煩わされることなく，言語現象を記述できるよう考慮されている。

最後に機械翻訳プロジェクトの成功には，翻訳者，コンピュータの専門家，言語学者それに実際の技術者の有機的な連係が必須であると考えられており，それがＴＡＵＭの推進に当っても反映されている。

(2) ＴＡＵＭ－ＡＶＩＡＴＩＯＮの概要[*1, *2]

ＴＡＵＭ－ＡＶＩＡＴＩＯＮは，英文の航空機マニュアル（水力学の章）をフランス語に翻訳する完全自動システムである。一旦入力された文章は，中途での人の介在することなく翻訳される。辞書での未登録語などで処理に失敗すると，何も結果は出さない。これは実用上厳しすぎると思われるが，その救済は実際のシステムのインプリメンテーションに当って対応すればよいとの考えのようである。

ＴＡＵＭ－ＡＶＩＡＴＩＯＮは解析，移行，生成の３フェーズからなり，その各々が更にいくつかのサブフェーズから構成されている。

---

*1 Bourbeau, Laurent 他: LINGVISTIC DOCUMENTATION OF THE COMPUTERIZED TRANSLATION CHAIN OF THE TAUM-AVIATION SYSTEM (1981).

*2 その他の文献については，Bourbeau, Laurent : DOCUMENTATION HIERARCHISÉE DU SYSTÈME TAUM/AVIATION (1980).

その軸となるのは，解析フェーズの結果であり，移行フェーズが操作する対象であり，生成フェーズの入力である文の中間表現で，正規表現と呼ばれるものである。正規表現は，本構造で表現され，述語の格支配構造を基本としている。これは，述語に対する選択制限を基にした意味処理との関係が強い要因となっているようである。選択制限は，動詞に対してだけでなく，形容詞に対しても行われる。前置修飾の範囲などに関しても，例えば日本語と英語などでは余り厳密に決定しない場合が多いが，フランス語では，後置修飾が多いこと，何よりも修飾するものとされるものの性数の一致が必要なことから，必須な条件となっている。

正規表現は，時制，性数などの情報が各ノードの属性値として吸い上げてしまった形であるのみならず，態やある種の文体の違い（“It is …”，間接目的の用法，形容詞の限定的用法と形容補語など）を，それらが変形操作によって等価となるものであるとき，同じ構造で表現することとしている。

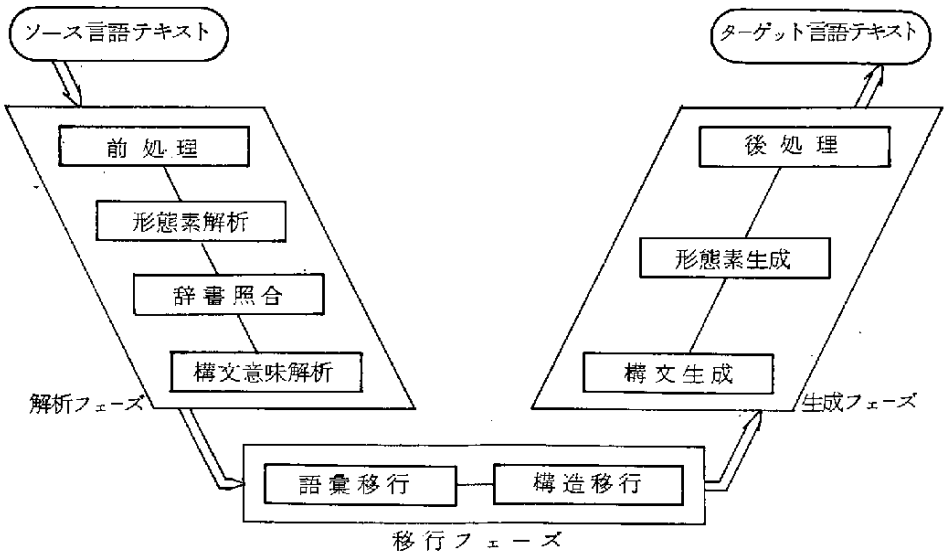


図3.4－1　TAUM－AVIATIONの処理フェーズ

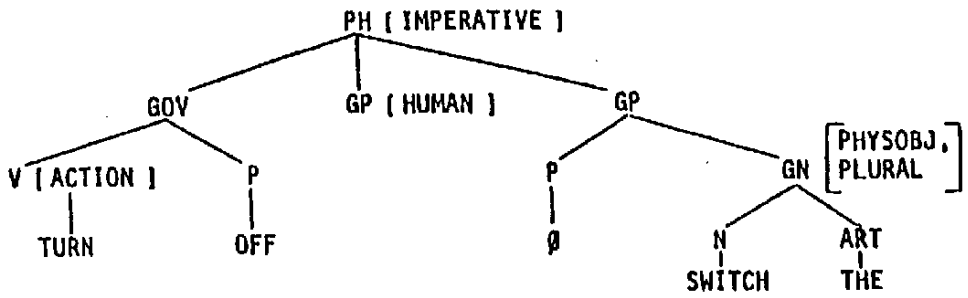


図 3.4－2　TAUM－AVIATIONの中間表現（正規表現）

TAUM－AVIATIONで用いている辞書は，解析フェーズで引用する解析用辞書と，語彙移行に際して用いられる移行用辞書とから成る。各々のエントリ数は，１９８１年３月時点で，解析用辞書が4,０５４，移行用辞書が3,２８０である。なお，各エントリは活用などにより屈折していない形態をとっている。これらの辞書記述は，言語学的にも相当に深いレベルまでの内容と正規表現に関する理解を必要とするので，少なくとも３カ月程度の特別な訓練を受けた翻訳や言語の専門家が行う必要があるとのことであった。

(3)　TAUM－AVIATIONの各フェーズ

TAUM－AVIATIONの処理単位は，原則として文である。ここには，通常の文の他に標題や図表の説明なども入る。したがって，名詞句，前置詞句なども一つの文単位となりうる。文にまたがる処理は，このシステムでは行っていない。

(a)　前　処　理

まず，語単位への分割が行われる。このとき，ピリオド，コンマ，括弧などの分離や略語のピリオドの認定などが必要となる。

(b)　形態素解析

形態素解析では，屈折形から原形と時制や数の推定を行う。その

LEXICAL CATEGORY	TOTAL
Nouns	1674
Adjectives	871
Verbs	833
Adverbs	187
Prepositions	149
Quantifiers	79
Ordinals	36
Pronouns	33
Subordinate conjunctions	28
Coordinate conjunctions	12
Articles	15
Equivalences rules (allographs)	136
TOTAL	4054

図 3.4－3　解析用辞書のエントリ数

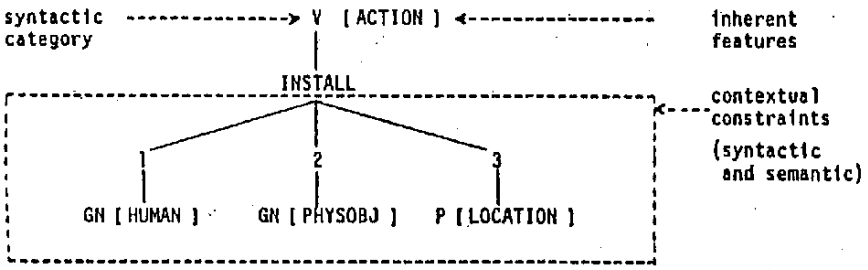


図 3.4－4　解析用辞書の語彙記述

結果，各語へいくつかの可能性が割り当てられる。

(c)　辞書照合

解析用辞書に対し，形態素解析で得られたいくつかのLexical

Word fromで照合を行い，辞書や語彙構造により置換される。

解析用辞書の各エントリは次の属性をもつ：

㈠品詞分類：名詞，動詞といった統語論的カテゴリ分類

㈡内在素性：「物質名詞」である／ないといった統語論的内在素性と「物体」，「流体」，「動作」といった意味論的内在素性。

㈢文脈素性：動詞文型などに代表される統語論的な文脈素性（選択制限情報）と，主語は人間である云々の意味論的文脈素性。

ここで，内在素性の設定は，サブランゲージの世界に依存しており，一般的な素性の他に，航空機マニュアルに独自の，そしてそこで強力なものを積極的に導入している。また，選択制限が比較的容易に設定できるのもサブランゲージへの限定の結果であろう。

(d) 構文意味解析

ＴＡＵＭ－ＡＶＩＡＴＩＯＮのパーサは，フィル・ウッズのＡＴＮ（Augumented Transition Network）を基として作られている。パージング自体は通常の文脈自由文法によっているが，同時に解析用辞書で指定された文脈素性と内在素性による選択制限によって，主に意味論的にアンビギュイティの解消を行っている。

この結果，英文に対する正規表現が作られる。

(e) 語彙移行

このシステムでは，語彙移行のフェーズでほとんどの英仏の対応がとられる。言い換えると，移行用辞書記述で非常に複雑な移行プロセスまでが記述されている。単純なものについては，英語の単語に対して，フランス語の単語を対応させればよいのであるが，他の語との関連で訳語を選択する必要も多くの場合出てくる。また，しばしば英仏対応において，構造的な変形も必要となってくる。そこ

動詞"check"の移行辞書記述：

```
if path cw ↑GOV ↑PH ↓GP〔3〕 ref arg3↓P ref arg3prep then if word
(arg3prep) is:
   #AGAINST#
      thendo begin
                translate cw as #COMPARER#
                translate arg3prep as #A#
             end;
   #FOR#
      thendo
      if path arg3(GN ref noungrp)
                 ↔GP〔2〕 ref arg2 ↓P ref arg2prep then begin
                translate cw as #VERIFIER#
                translate arg2prep as #DE#
                translate arg3prep as #ø#
                move arg2 into GP position under noungrp end
      end
   else translate cw as #VERIFIER#.
```

（意味）

"checkXagainstY"なら"comparerXàY"に

"checkXforY"なら"vérifierYdeX"に

それ以外なら単に"check"を"vérifier"に翻訳する。

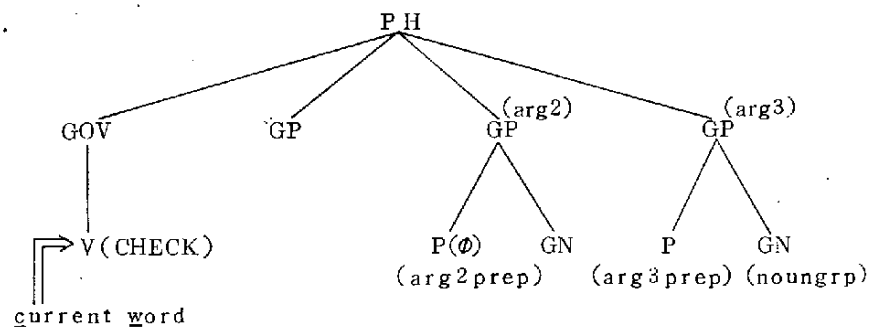


図3.4－5　移行用辞書の記述例

で移行用辞書は，移行のプロセスをも記述した一種の目的向け言語により記述する。語彙移行のフェーズは，この移行用辞書の語彙記述に従って処理を行うプロセッサである。

(f) 構造移行

TAUM－AVIATIONの構造移行フェーズは，語彙レベルではあらかじめ指定のできない文全体に関する変換だけを行う。具体的には，時制の一致に関する英語とフランス語での扱いの違いなどがある。例えば，フランス語では，主文が未来時制のとき従属文に関しても，英語の場合と異なり未来時制が要求される。

(g) 構文生成

フランス語の構造へ移行された正規表現を，性・数・時制などの属性を伴った語の列へと分解する。こうした属性情報は，性数の一致の規則を用いても付与される。

(h) 形態素生成

構文生成で得られた各語の語彙と属性情報から，フランス語のMorphological Word form を作る。また，母音の省略や縮約などもこのフェーズで行われる。

(i) 後 処 理

実際のフランス語テキストが作られる。

形態素解析と形態素生成を除いて，各フェーズ用の専用言語が用意され，各フェーズはそれを用いて記述されている。また，プログラミング言語は，構造移行と構文生成のフェーズの為の言語Q－SYSTEMSがFORTRAN で作られている他は，PASCALが用いられている。

(3) TAUM－AVIATIONの評価

TAUMの資料によると，TAUM－AVIATIONの実働テストは未だ十分なものとはいえず，4万語程度が処理されたぐらいだと

いう。テストの評価によると，入力された文単位に対して3分の2程が処理され，翻訳者による翻訳結果に対して８０％程度が適訳であった。また，ＴＡＵＭ－ＡＶＩＡＴＩＯＮに対して与えられた翻訳コストは，研究開発に要した費用は除外して，次の様な数字が示されている。

TAUM−AVIATIONを用いたコスト	0.１８３ドル／語
人間によるコスト	0.１４５ドル／語

TAUM−AVIATIONを用いたコストの内訳：

準備及び入力作業	８％
自動翻訳処理	４３％
人間の後編集	３７％
転記と校閲	１２％

自動翻訳処理費用の内訳：

ＣＰＵ処理費用	0.０４６８ドル／語
入出力料金	0.００３２ドル／語
メモリ使用	0.０３３１ドル／語

フェーズ毎の割合では，次の様になっている：

前処理	2.１％
形態素解析	3.１％
辞書照合	１９.１％
構文意味解析	３４.２％
語彙移行	１３.０％
構造移行と構文生成	２０.３％
形態素生成	1.１％
後処理	7.１％

テストはモントリオール大学計算センターのＣＤＣ１７３，NOSBE 1.4，ＬＥＶＥＬ５０８のＯＳを用いて行われた。処理時間は一語当

たりＣＰＵ時間0.４２秒，入出力時間0.２１秒であったという。

これらの評価はあくまで研究開発中のシステムを用いて行われたものであり，実際のインプリメントの状態とは相当違うと思われるので，この結果をそのまま最終的な性能評価と結びつけるのは早計とは思われるが，傾向は見えて興味深いものである。

ＴＡＵＭのシステムは，限定した範囲を対象としているといいながら，非常にオーソドックスな翻訳システムであり，モジュール化なども積極的に行われている，機械翻訳システムのモデルともいうべきシステムである。辞書記述などで，非常に煩雑と思われるような部分もあるが，それは余剰規則などを導入することにより相当程度は解消できるであろう。実現にあたっては，マンマシン・インターフェイスや言語学的な割り切りをはじめとする，アドホックな種々の要因が加わることになるであろうが，マクロには，いまだに機械翻訳システムの範となるようなシステムであることは変わっていないと思われる。

3.4.2　文章情報データベースの現状調査

《その１》ＩＳＩ社

ＩＳＩ社では，同社の科学技術文献データベースサービスと自動インデックスの現状について調査した。

(1)　ＳＣＩＳＥＡＲＣＨ（ Science Citation Index SEARCH ）

学術文献には，参照文献を示すのが通例である。論文の最後に載せられている参照（引用）文献リストやテキスト中での引用を拾い出してパンチし，サイテーション・インデックスという索引を作り，これをデータベースに収録したものである。

１９６０年から今日までの科学技術文献に関する非常に大型の索引データベースとなっている。業界雑誌，学界雑誌など約1,３００誌から約5,０００タイトルを包含している。これらは毎年増大しているが，

ちなみに１９８２年は７００,０００件の増加であり，この内５００,０００件が自然科学，残りが社会科学，人文科学に関するものである。後者のデータベースは，ＳＳＣＩＳＥＡＲＣＨ（ Social Sciences Citation Index SEARCH ）と呼ばれている。これらは，１年に１回，コンピュータ出力のハードコピーで提供され，また５年に１度まとめて出版されている。さらにロッキード社のDialog システムを通じてオンラインで利用できるようになっている。

サイテーション・インデックスは，専任のエディタが論文からそのタイトル，著者名，住所，そこで引用されている論文をすべてデータエントリできるようにしたものである。科学分野の場合には，引用文献はページの下方に脚注の形で示すか，論文の最後に列記されているので比較的容易に抽出できる。特に人文科学等では，引用文献がテキスト中に織り込まれていることがあるので，これらもテキストにざっと目を通して拾い出している。

利用者はサイテーション・インデックスによって，たとえその言葉が文献に出現していなくても，何がトピックスになっているかを，科学のあらゆる分野に，互いに相関関係をもって調べることができる。例えば，１つの事柄が２つの分野にまたがっているときでも検索できるような仕組みになっている。

ＩＳＩ社では，この他に次のデータ提供サービスを行っている。

・ＡＳＣＡ（ Automatic Subject Citation Alert ）レポート

これはＳＤＩ（ Selective Dissemination of Information ）システムで，週ごとの色々な情報のトピックスについてユーザ・プロファイルに該当するものを提供するサービスである。

・Current Abstruct of Chemistry

有機化学に関して，最近の学術雑誌からの抜粋版である。これはすべてグラフィックで表わされている。化学構造式から構造図を描

き，１つ１つの構造式が反応によってどのように変っていくかを図式で表わしている。

・化学化合物に関するデータベース

・有機化学の反応に関するデータベース

・パーソナル・データ・マネージャ“ＳＩＭＡＴＥ”

これはパーソナルコンピュータから，同社のデータベースを検索して，結果をパーソナルコンピュータ側にダウン・ロードすることができるシステムである。パーソナルコンピュータは，ＩＢＭ，アップル，ＴＲＳなどが使える。

(2) Atlas of Science

１９７８年と１９８０年のＳＣＩの生化学及び分子生物学のデータから作ったものである。これは，クラスタ分析による情報に基づいて１２０分類にわたって，どのような文献が最もよく引用され，文献どおしがどのような相互関係をもっているかを地図的に目で見ることができるようになっている。

これは次の４つの部分より構成されている。

① ミニレビュー（図3.4－6参照）

それぞれのクラスタ・マップで表わされているトピックスに関連した評論である。ユーザはこれによって容易にそのトピックスの概要を知ることができる。

② クラスタ・マップ（図3.4－7，3.4－8参照）

コア・ドキュメント（重要な発明・発見を述べた文献）間の関連性を地図的に表わしたものである。□内の数字はコア・ドキュメントを意味しており，クラスタ・マップの下方に掲載されているビブリオグラフィとの対応に使われる。マップ中の格子は分野の方向を表わし，表示位置が近ければ近い程，相関関係の強い分野どおしのドキュメントである。

図3.4－8は，コア・ドキュメント間の引用関係を付け加えたものである。ユーザはこれによって，どのような文献がどの分野において重要な役割をなしているか，また，どのような文献が相互に相関関係をもっているかを見ることができる。

③　コア・ドキュメントのビブリオグラフィ（図3.4－7，3.4－8参照）

マップ中のコア・ドキュメントの書誌的情報である。ＣＦの数字は他の文献から1年間に引用された回数を表わす。

④　コア・ドキュメントを引用している文献リスト（図3.4－9参照）

ＲＷの数字は，コア・ドキュメントに対する参照回数によるウェート付けを表わす。また最近の研究についてだけは，特別にリストされていて，その分野の進展状態が分るようになっている。

このようにAtlas of Science は，科学の分野を地図で表わし，目で見れるようにしたものであるが，今後対象文野を広げていって，

Nitrogen-Fixation by Rhizobia

Nitrogen (N_2) fixation is accomplished by few bacterial groups. other organisms are dependent on a source of fixed nitrogen to meet their metabolic requirements, either in an oxidized form (e.g., nitrate), or a reduced form such as ammonium sulfate or amino acids. The key reaction of N_2 fixation is carried out by nitrogenase, a complex enzyme consisting of at least two protein subunits, that reduces N_2 to the level of NH_3. This requires energy supplied as ATP, and a strong reductant such as ferredoxin. A characteristic of nitrogenase is that it will reduce acetylene (C_2H_2) and this reaction is the basis of a widely used assay system for the enzyme.

The Rhizobia fix N_2 symbiotically with plant hosts belonging to the legume family. The bacteria invade the root cells and form nodules, providing the host plant with nitrogen that can be assimilated in return for a source of carbohydrate. The precise nature of the association has interested researchers for many years, especially because the bacteria have a very beneficial effect on crop production. Until recently, it was not possible to induce Rhizobia to fix N_2 without their natural hosts. Indeed, the association appeared to be so obligatory that it was suggested the plant may be supplying genetic factors necessary for the development of nitrogenase. However, it now appears that it was just a matter of getting growth conditions right for the bacteria to produce nitogenase and fix N_2 in culture.

In the early 1970's, there were reports that Rhizobia would fix N_2 in the presence of cultured cells of the host plant or other legumes closely related to the natural host. Then Child [1] showed that *Rhizobium* cowpea strain 32 H1 was able to fix N_2 when grown on agar with plant cell callus cultures from several legumes, and also three nonlegumes—rape, wheat, and brome grass. The bacterial colonies were observed by microscopy to be free-living on the surface of the callus and between the cells. If the callus tissue was removed, bacteria remaining on the agar were also able to fix N_2 to a limited extent. These results suggested that some diffusible factor(s) from the plant cells were required for nitrogenase to be expressed. Similar findings were published simultaneously by Scowcroft and Gibson.[1]

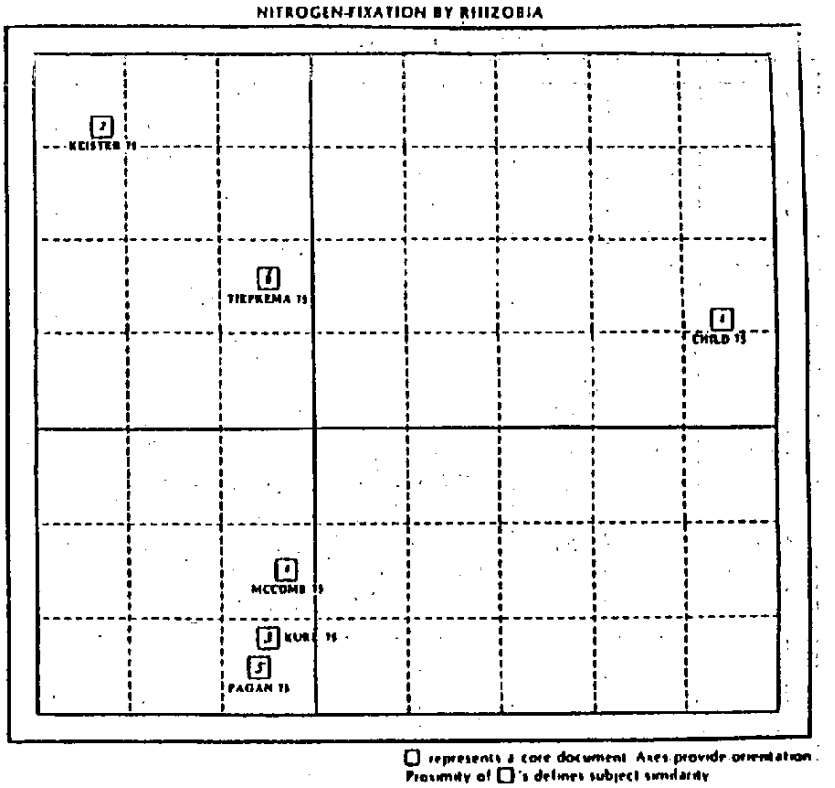
The race was now on to find out what the diffusible factors were. Later in the same year five papers were published, three of them in the same issue of *Nature*, describing culture media and conditions under which Rhizobia would fix N_2 without plant cells. Pagan et al. [2] found that the crucial components for an agar based medium were a sugar (arabinose, galactose), a tricarboxylic acid cycle intermediate (succinate, fumarate), and, perhaps unexpectedly, a source of 'ready-fixed' nitrogen (ammonium sulfate, glutamine). As usual, nitrogenase activity was assayed by C_2H_2 reduction, but direct incorporation of N_2 was checked by culturing the bacteria in an atmosphere containing the heavy isotope N^{15}. Similar, if not identical, results were reported by Kurz and La Rue [3] and McComb et al. [4]. In addition, Tjepkema and Evans [5] and Keister [6] found that low concentrations of oxygen were required for optimum nitrogenase activity in a liquid culture medium.

These results have established that the genes necessary to code for functional nitrogenase are present in Rhizobia. Not all strains of *Rhizobium* tested have been observed to fix N_2 under these improved culture conditions, but the list is increasing [3]. The requirements of some Rhizobia are very exacting, as is evident by their host plant specificity. Additional factors and different levels of nutrients may be necessary to induce some strains to fix N_2 in culture.

1986 Supplement

Recent studies indicate that the genes involved in nitrogenase expression, as well as those for several other symbiotic functions, are located in plasmids rather than chromosomes [58]. Glutamate synthetase appears to play a crucial role in regulating nitrogenase expression. In Rhizobia, glutamate synthetase occurs in two forms, one of which can have its catalytic activity modulated by reversible adenylylation. Experiments in culture with glutamine auxotrophs and their revertants suggest that only this form of the enzyme is concerned with nitrogenase de-repression [510].

図3.4－6　ミニレビューの例

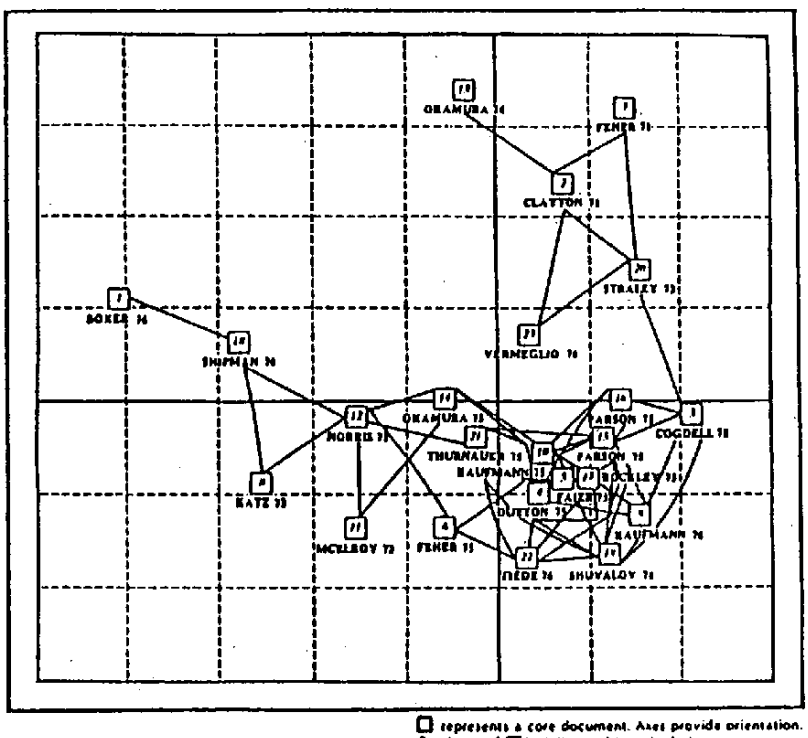


Cited Core Documents

		CF
1	CHILD JJ Nitrogen-fixation by a Rhizobium SP in association with non-leguminous plant-cell cultures *Nature* 253(5490) 350, 1975	18
2	KEISTER DL Acetylene-reduction by pure cultures of Rhizobia *J Bacteriol* 131(3) 1265-1268, 1975 N	19
3	KURZ WGW, LARUE TA Nitrogenase activity in Rhizobia in absence of plant host *Nature* 256(5516) 407-409, 1975	30
4	MCCOMB JA, ELLIOTT J, DILWORTH MJ Acetylene-reduction by Rhizobium in pure culture *Nature* 256(5516) 409-410, 1975	26
5	PAGAN JD, CHILD JJ, SCOWCROF WR, GIBSON AH Nitrogen-fixation by Rizobrum cultured on a defined medium *Nature* 256(5516) 406-407, 1975	28
6	TIEPKEMA J, EVANS HJ Nitrogen-fixation by free-living Rhizobium in a defined liquid medium *Biochem Biophys Res Commun* 65(2) 625-631, 1975	22

図3.4－7　コア・ドキュメントのクラスタ・マップとビブリオグラフィの例(1)

科学情報の百科辞典として収録していきたいとのことである。これによって，非常に広範で複雑な科学の分野を，簡単に，一見してその関連を見ることができるようになるであろうとのことである。



□ represents a core document. Axes provide orientation. Proximity of □'s defines subject similarity

Cited Core Documents

[1] BOXER SG, CLOSS GL — CF 19
Covalently bound dimeric derivative of pyrochlorophyllide A—possible model for reaction center chlorophyll
J Am Chem S 98(17):5406-5408, 1976 L

[2] CLAYTON RK, WANG RT — CF 22
Photochemical reaction centers from Rhodopseudomonas-spheroides
Meth Enzym 23:696, 1971

[3] COGDELL RJ, MONGER TG, PARSON WW — CF 21
Carotenoid triplet-states in reaction centers from Rhodopseudomonas-sphaeroides and Rhodospirillum-rubrum
Bioc Biop A 408(3):189-199, 1975

[4] DUTTON PL, KAUFMANN KJ, CHANCE B, RENTZEPI PM — CF 35
Picosecond kinetics of 1250 NM band of RPS-sphaeroides reaction center—nature of primary photochemical intermediary state
FEBS Letter 60(2):275-280, 1975

[5] FAJER J, BRUNE DC, DAVIS MS, FORMAN A, SPAULDIN. LD — CF 41
Primary charge separation in bacterial photosynthesis—oxidized chlorophylls and reduced pheophytin
P NAS US 72(12):4956-4960, 1975

[6] FEHER G, HOFF AJ, ISAACSON RA, ACKERSON LC — CF 24
Endor experiments on chlorophyll and bacteriochlorophyl invitro and in photosynthetic unit
Ann NY Acad 244(Apr15):239-259, 1975

[7] FEHER G — CF 23
Some chemical and physical properties of a bacterial reaction center particle and its primary photochemical reactants
Photochem P 14(3):373, 1971

[8] KATZ JJ, NORRIS JR — CF 35
Chlorophyll and light energy transduction in photosynthesis
Curr T Bio 5:43, 1973

[9] KAUFMANN KJ, PETTY KM, DUTTON PL, RENTZEPI PM — CF 17
Picosecond kinetics in reaction centers of RPS-sphaeroides and effects of ubiquinone extraction and reconstitution
Bioc Biop R 70(3):839-845, 1976

[10] KAUFMANN KJ, DUTTON PL, NETZEL TL, LEIGH JS, RENTZEPI PM — CF 39
Picosecond kinetics of events leading to reaction center bacteriochlorophyll oxidation
Science 188(4195):1301-1304, 1975

[11] MCELROY JD, FEHER G, MAUZERAL DC — CF 16
Characterization of primary reactants in bacterial photosynthesis .3. Comparison of light-induced EPR signal (g = 2.0026) with that of a bacteriochlorophyll radical
Bioc Biop A 267(2):363, 1972

[12] NORRIS JR, UPHAUS RA, CRESPI HL, KATZ J — CF 28
Electron-spin resonance of chlorophyll and origin of signal-I in photosynthesis
P NAS US 68(3):625, 1971

[13] OKAMURA MY, STEINER LA, FEHER G — CF 22
Characterization of reaction centers from photosynthetic bacteria .3. Subunit-structure of protein mediating primary photochemistry in Rhodopseudomonas-spheroides R-26
Biochem 13(7):1394, 1974

[14] OKAMURA MY, ISAACSON RA, FEHER G — CF 24
Primary acceptor in bacterial photosynthesis—obligatory role of ubiquinone in photoactive reaction centers of Rhodopseudomonas-spheroides
P NAS US 72(9):3491-3495, 1975

図 3.4－8　コア・ドキュメントのクラスタ・マップと

ビブリオグラフィの例(2)

Key Citing Documents

		RW
1	DILWORTH MJ, MCCOMB JA Recent advances in tissue-culture studies of Legume-Rhizobium symbiosis (Ayanaba A, Dart PJ, eds) *Biological Nitrogen Fixation in Farming Systems of the Tropics* New York: John Wiley and Sons Inc, 1977 p 135	6
2	GIBSON AH, PAGAN JD, SCOWCROF WR Nitrogen-fixation in plants—expanding horizon (Newton W, Postgate JR, Rodriguez-barrueco C, eds) *Recent Developments in Nitrogen Fixation* London: Academic Press, 1977 p 307	6
3	LORKIEWI Z, RUSSA R, URBANIK T Nitrogen-fixation by Rhizobium in pure cultures *Act Micro P* 27 S, 1978	6
4	WILCOCKS J, WERNER D Nitrogenase activity of Rhizobium-japonicum growing on agar surfaces in relation to slime production, growth and survival *J Gen Micro* 108:151, 1978	6
5	BERGERSE FJ Nitrogenase in chemostat cultures of Rhizobia (Newton W, Postgate JR, Rodriguez-barrueco C, eds) *Recent Developments in Nitrogen Fixation* London: Academic Press, 1977 p 309	5
6	BERGERSE FJ Factors controlling nitrogen-fixation by Rhizobia (Ayanaba A, Dart PJ, eds) *Biological Nitrogen Fixation in Farming Systems of the Tropics* New York: John Wiley and Sons Inc, 1977 p 133	5
7	KANESHIR J, CROWELL CD, HANRAHAN RF Acetylene-reduction activity in free-living cultures of Rhizobia *Int J Sy B* 26:27, 1976	5
8	KEISTER DL, RAO VR Physiology of acetylene-reduction in pure cultures of Rhizobia (Newton W, Postgate JR, Rodriguez-barrueco C, eds) *Recent Developments in Nitrogen Fixation* London: Academic Press, 1977 p 419	5
9	PANKHURS CE, CRAIG AS Effect of oxygen concentration, temperature and combined nitrogen on morphology and nitrogenase activity of Rhizobium-SP-strain 32H1 on agar culture *J Gen Micro* 106:207, 1978	5
10	SADANA JC, KHAN BM Nitrogen-fixation *J Sci Ind R* 36:510, 1977 R	5
11	SHANMUGA KT, ANDERSEN K, OGARA F, VALENTIN RC Biological nitrogen-fixation *Ann R Plant* 29:263, 1978 R	5
12	SHANMUGA KT, ANDERSEN K, MORANDI C, OGARA F, VALENTIN RC Genetic-control of nitrogen-fixation (NIF) (Newton W, Postgate JR, Rodriguez-barrueco C, eds) *Recent Development in Nitrogen Fixation* London: Academic Press, 1977 p 321	5
13	SKOTNICK ML, ROLFE BG Differential stimulation and inhibition of growth of Rhizobium-Trifolii strain 31 and other Rhizobium species by various carbon-sources *Microbios* 20:15, 1977	5
14	UPCHURCH RG, ELKAN GH Ammonia assimilation in Rhizobium-japonicum colonial derivatives differing in nitrogen-fixing efficiency *J Gen Micro* 104:219, 1978	5
15	YATES MG Physiological aspects of nitrogen-fixation (Newton W, Postgate JR, Rodriguez-barrueco C, eds) *Recent Developments in Nitrogen Fixation* London: Academic Press, 1977 p 219	5

Supplementary Citing Documents

		RW
1	KURZ WGW, CHILD JJ Asymbiotic fixation of dinitrogen by Rhizobium in vitro (Sanpietro A, ed) *Photosynthesis and Nitrogen Fixation* Pt C, New York: Academic Press, 1980 p 750	6
2	DAVEY MR, COCKING EC, PEARCE N Fusion of legume root nodule protoplasts with non-legume protoplasts—ultrastructural evidence for the functional-activity of Rhizobium bacteroids in a heterokaryotic cytoplasm *Z Pflanzenp* 99:415, 1980	5
3	GILES KL, VASIL IK Nitrogen-fixation and plant-tissue culture (Vasil IK, ed) *Perspectives in Plant Cell and Tissue Culture* Pt B, New York: Academic Pres, 1980 p 81 R	5
4	ROBSON RL, POSTGATE JR Oxygen and hydrogen in biological nitrogen-fixation *Ann R Micro* 34:183, 1980 R	5
5	SEN D, SCHULMAN HM Enzymes of ammonia assimilation in the cytosol of developing soybean root-nodules *New Phytol* 85:241, 1980	5
6	TJEPKEMA JD, ORMEROD W, TORREY JG Vesicle formation and acetylene-reduction activity in Frankia SP CP II cultured in defined nutrient media *Nature* 287:633, 1980	5
7	VANBERKU P, BOHLOOL BB Evaluation of nitrogen-fixation by bacteria in association with roots of tropical grasses *Microbiol R* 44:491, 1980 R	5
8	BERINGER JE, BREWIN NJ, JOHNSTON AW The genetic-analysis of Rhizobium in relation to symbiotic nitrogen-fixation *Heredity* 45:161, 1980 R	5
9	BRILL WJ Nitrogen-fixation (Carlson PS, ed) *Biology of Crop Productivity*, New York: Academic Press, 1980 p 51	5
10	LUDWIG RA Regulation of Rhizobium nitrogen-fixation by the uridylylated glutamine synthetase-II system *P Nas Acd* 77:1617, 1980	5

図 3.4－9　コア・ドキュメントを引用している
ドキュメントのリスト例

(3) 自動インデックス

ＩＳＩ社では，次の２つの自動インデックス手法が研究・開発されている。

① ＰＳＩ社（ Permuterm Subject Index ）

ＰＳＩは，full-stop list と semi-stop list を使って，文献タイトルから重要語のペア・リストを作って索引とするものである。（図3.4－6参照）

ここで，full-stop list とは前置詞，接続詞，冠詞などを除外するものであり，semi-stop list とは“method”などの語を除外するものであるが，従属語として出現した場合には有効とするものである。タイトル中での重要語のペア・リストは，可能な組み合せをすべて表示するのでどうしても多くなりがちである。これを防止するために，２～３語のフレーズについて統計的に共出現するも

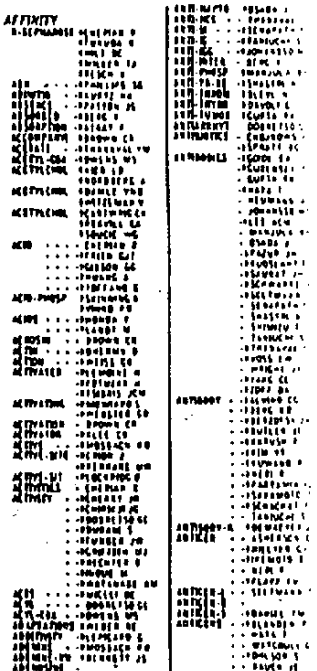
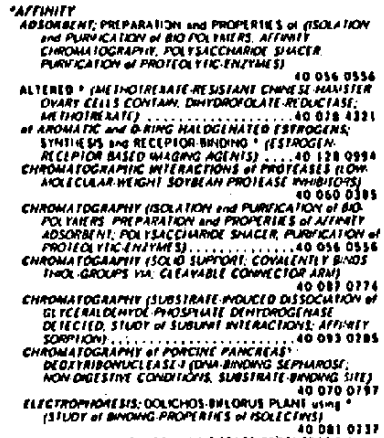


図3.4－6　ＰＳＩの例



*AFFINITY
ADSORBENT; PREPARATION and PROPERTIES of (ISOLATION and PURIFICATION of BIO POLYMERS. AFFINITY CHROMATOGRAPHY. POLYSACCHARIDE SHACER PURIFICATION of PROTEOLYTIC-ENZYMES)
40 056 0556

ALTERED * (METHOTREXATE-RESISTANT CHINESE HAMSTER OVARY CELLS CONTAIN. DIHYDROFOLATE-REDUCTASE; METHOTREXATE) 40 078 4321
of AROMATIC and O-RING HALOGENATED ESTROGENS; SYNTHESIS and RECEPTOR-BINDING * (ESTROGEN-RECEPTOR BASED IMAGING AGENTS) 40 128 0994
CHROMATOGRAPHIC INTERACTIONS of PROTEASES (LOW-MOLECULAR-WEIGHT SOYBEAN PROTEASE INHIBITORS)
40 060 0385
CHROMATOGRAPHY (ISOLATION and PURIFICATION of BIO-POLYMERS PREPARATION and PROPERTIES of AFFINITY ADSORBENT; POLYSACCHARIDE SHACER, PURIFICATION of PROTEOLYTIC-ENZYMES) 40 056 0556
CHROMATOGRAPHY (SOLID SUPPORT, COVALENTLY BINDS THIOL-GROUPS VIA, CLEAVABLE CONNECTOR ARM)
40 083 0776
CHROMATOGRAPHY (SUBSTRATE-INDUCED DISSOCIATION of GLYCERALDEHYDE-PHOSPHATE DEHYDROGENASE DETECTED; STUDY of SUBUNIT INTERACTIONS; AFFINITY SORPTION) 40 093 0285
CHROMATOGRAPHY of PORCINE PANCREAS' DEOXYRIBONUCLEASE-I (DNA-BINDING SEPHAROSE; NON-DIGESTIVE CONDITIONS, SUBSTRATE-BINDING SITE)
40 070 0797
ELECTROPHORESIS; OOLICHOS-BULORUS PLANT using * (STUDY of BINDING-PROPERTIES of ISOLECTINS)
40 081 0337
of HEMOGLOBIN; OXYGEN MULTISTAGE REGENERATION PROCESS and INVIVO REDUCTION of LIPOSOMES of ALLOSTERIC EFFECTORS; OXYGEN) 40 120 0502
LABEL; NLW * (ADENOSINE 5'-(2-BROMOETHYL)-PHOSPHATE; ADENINE-NUCLEOTIDE SITES; PROTEINS)
40 072 7517
of SIDE-CHAIN HALOGENATED HEXESTROL DERIVATIVES; SYNTHESIS and RECEPTOR-BINDING * (ESTROGEN-RECEPTOR BASED IMAGING AGENTS) 40 128 1002
SORPTION (SUBSTRATE-INDUCED DISSOCIATION of GLYCERALDEHYDE-PHOSPHATE DEHYDROGENASE DETECTED; AFFINITY CHROMATOGRAPHY; STUDY of SUBUNIT INTERACTIONS) 40 093 0285

*AFFINITY-PURIFIED
GUANINE-NUCLEOTIDE REGULATORY PROTEIN (RESTORATION of GUANINE NUCLEOTIDE-STIMULATED and FLUORIDE-STIMULATED ACTIVITY; ADENYLATE CYCLASE-DEFICIENT CELL-LINE) 40 061 0439

図3.4－7　KWPＳＩの例

のは，－（ハイフン）で結合して一語として扱っている。

② KWPSI（Key Word／Phrase Subject Index）

伝統的なインデクシングは，通常，語を索引としていたが，これはタイトルを文節に分けて，それを索引とするものである（図3.4－7参照）。科学の分野では，毎年34,000～40,000語位の新しい言葉が出てきている。従って，語だけを抽出して辞書に入れていく方法では，とてもそれに追従していくことはできないので，構文解析によってタイトルをフレーズに区切って索引とする方法を開発した。タイトルは構文的にシンプルであるので構文解析も容易に実現できた。

現在，オンライン検索では，PSIの索引が用いられているが，KWPSIの方がより文献内容を表わしており，検索に有効である。今後は，オンラインでのフレーズ検索をめざしているとのことである。

《その2》UCLA

UCLAでは，Harold Borko 教授を訪ね文章情報データベースの現状及び問題点について教えていただいた。

(1) データベースの統合化

専門分野のデータベースは，分野別にできているのが現状である。このように別個にもった場合，利用者側に問題が生じる。

例えば，医学の場合，特定の薬の使用に関心のある者にとって，化学データベース，医学データベース，さらに生物学データベースなどを調べなければならない。このときデータベースごとに異なる検索をすることになる。これを避けるためには，利用者の質問文は1つで，これに関連して色んなデータベースから自動的に応答してくれる仕掛けが必要になってくる。これはシソーラスの統合化でなく，単一質問

からデータベースごとに索引語を自動的に変換してくれる方法が望ましい。

(2) 自動インデックスの現状

ここ数年間，あまり変化はないようである。自動インデックスには，2つの方法がある。

その1つは，コンピュータを使ってシソーラスを作るが，判断は人間が行うRelevence-Feedback 法である。これは，データベースはインデックスされてないと仮定して，人間が質問やニーズを表わす。するとコンピュータはその質問の中における色々と関連性のある言葉を探し出し，これを使ってデータベース中のいくつかの文献を探し出す。スクリーン上に2つずつ文献のアブストラクトを人間に見せ，どれが最も関係のある文献であるかを問う。人間が関連性のある文献のウェイトを高く付けて，同様に残りの文献についても関連付けをしていく方法である。

われわれは，人間のために検索をやっているのであるから，システムの中で人間がこの種の判断をくだすのは良い方法だと思われる。

もう1つの方法としては，構文解析による方法であろう。これについては，現時点ではわからないが，数年前に名詞句を識別できる研究が行われた。英語の場合であるが，索引語の90%が文献タイトル中やパラグラフの最初のセンテンス中に現われてくるといわれている。この方法で自動インデックスを作成する研究がなされた。

それ以降ある程度の進展はあったと思うが，フルテキストを構文解析する方法は経費がかかり，結果的なメリットが少ない。

ただし，ここ数年次の2つの理由で環境は変りつつあるといえる。

- 文章情報は，コンピュータ出力されたものが増えたこと（すなわち，マシン・リーダブルである）。
- コンピュータのメモリ，スピード，コストパフォーマンスが飛躍

的に向上したこと。

このことから，構文解析による自動インデックスもコスト面で有意義なものとなろう。

(3) 用語管理について

用語管理に関するレポートとしては，ＥＲＩＣ（Educational Resources Information Center）システムのものがある。これは米国政府の教育用データベースであるが，教育におけるボキャブラリはそれ程基準化されていないところから，シソーラスのメンテナンスが非常に難しい。そのために，シソーラスをメンテナンスするプログラムが開発された。

このようなプログラムは，完全自動で行うのでなく，コンピュータと人間のコンビネーションで行うのが良い。ある言葉がデータベースの中で何回出現するかのカウントはコンピュータでできる。例えば，５０,０００に達する文献の中で１～２回しか出現しない言葉であれば索引語から除外してよい。同様に，２０,０００回も出現しているものも除外できる。また，ユーザが質問文の中である索引語を何回使ったかもコンピュータが教えてくれることができる。ある索引語が正しい文献を検索するかどうかは人間が判断を下すしかない。従って，コンピュータが抽出した言葉をシソーラスに入れるのは人間が行うことになる。

シソーラスは常にメンテナンスしておかないと，正確な検索ができなくなる。新しい用語を付け加えていく一つの方法がある。ある文献に新しい用語（複合語の場合が多い）が出現した場合には，即シソーラスに登録せず，いったん未定義語として別のファイルに管理しておく。そして，例えば１０個以上使われたら，はじめてシソーラスに登録する。この場合も，シソーラスに入れるかどうかの判断は人間が行うべきである。

そして，ボキャブラリは年々変化していくので，数年に１回はシソーラス全体の見直しをやるべきである。

3.5 ＰＲＯＬＯＧ言語を利用した文章解析処理の実験

3.5.1 背　　景

最近第５世代コンピュータにおいて中心的な問題向きの処理言語となるであろうと予想されているPROLOG[1),2)]にとみに関心が高まっている。ＰＲＯＬＯＧという名称は，Programming in logicに由来し，その名のとおり一階述語論理に基づくプログラミング言語である。ＰＲＯＬＯＧ自体は，英ロンドン大学のKowalski[1)]らのグループの考案になるもので，１９７２年に仏マルセイユ大学のColmerauerのグループがコンピュータ上にインプリメントしたのが最初とされている。[3)]

その後エジンバラ大学等でも改良が加えられているが，我が国でも，メーカ，大学，研究機関でその研究開発がなされている。

従来のプログラミング言語が，ハードウェアの動作に符合した手続き中心型であるのに対し，ＰＲＯＬＯＧでは手続きはなく，ただデータもプログラムもホーン節（一階述語論理式）の集合で記述される論理定義型である。

従って，プログラムそのものには，何をしたいかをのみ記述すればよく，従来のプログラミング言語では当り前のこととされていた処理のための処理を必要としない。

我々は，単純ではあるが種々の点に可能性を持つＰＲＯＬＯＧ言語がどの程度文章解析処理システムに対して適応性を持つかということを確認する目的で機械翻訳の実験を行った。以下実験について述べたものである。

3.5.2　ＰＲＯＬＯＧとデータベース

ＰＲＯＬＯＧの記述能力は次の２つに大別される。

- 規則の記述
- 事実の記述

(1)　規則の記述

規則とは，２つまたは，それ以上の事柄の間に論理的な因果関係を持つことをいい，次のような表現をいう。

「もし～であるならば～である」

これをＰＲＯＬＯＧで一般的に表現すると次のようになる。

$P :- Q_1 [, Q_2 \cdots]$.

：－は，「もし～ならば」の意味を持つ。次にＰＲＯＬＯＧにおける規則の記述例を示す。

likes（ＪＯＨＮ，＊Ｘ）：－likes（ＴＯＭ，＊Ｘ）．

意味：もし，ＴＯＭがそれを好きならば，ＪＯＨＮもそれが好き。

likes（ＪＯＨＮ，＊Ｘ）：－ loves（ＪＯＨＮ，＊Ｙ），
likes（＊Ｙ，＊Ｘ）。

意味：もし，ＪＯＨＮが愛している人がそれを好きならば，ＪＯＨＮもそれが好き。

このように，Q_1，Q_2 ・・・ で示されるように，右辺のカンマで記した関係は，ＡＮＤ条件となり，それらの条件のすべてが満たされれば左辺が成立する。

もし，この定義を数行にまたがって記述すれば，ＯＲ条件となり，例えば次の意味をもつ。

likes（ＪＯＨＮ，＊Ｘ）：－likes（ＴＯＭ，＊Ｘ）．

likes（ＪＯＨＮ，＊Ｘ）：－likes（ＭＡＲＹ，＊Ｘ）．

意味：もし，ＴＯＭかまたはＭＡＲＹがそれを好きならば，ＪＯＨＮもそれが好き。

これらの規則は，手続きに関係なく（但し，定義した順序は，その規則の検証の順序になる）ただ定義するだけで，ＰＲＯＬＯＧが自動的に真否を検証してくれる。従ってプログラマは，不要な処理から解放され，もっぱら規則の正当性のチェックにのみ専念できる。また，この規則の定義は，後に述べる文章における文法規則を記述する上で極めて有用なものである。

(2) 事実の記述

事実とは，いくつかの事物の間にある関係が成立つことをいい，次のような表現をいう。

「～　は　～　である」

これをＰＲＯＬＯＧで一般的に表現すると，次のようになる。

Ｐ．

つまり，先に述べた規則より：－と，右辺のすべてをピリオドを残して取去った形である。結果的には，「もし～ならば」の表現がなくなっただけである。次にＰＲＯＬＯＧにおける事実の記述例を示す。

likes（ＪＯＨＮ，ＭＡＲＹ）
likes（ＴＯＭ，ＢＥＴＴＹ）
likes（ＪＩＭ，ＢＥＴＴＹ）

意味：ＪＯＨＮは，ＭＡＲＹが好きである。
　　ＴＯＭは，ＢＥＴＴＹが好きである。
　　ＪＩＭも，ＢＥＴＴＹが好きである。

この定義形式は，リレーショナル・データベース[4]におけるリレーションと同一のものであり，例えば，従業員リレーションをＰＲＯＬＯＧで記述すると次のようになる。

ＥＭＰ（ＪＯＨＮ，２５，＄５００）．
ＥＭＰ（ＴＯＭ，３１，＄７００）．
⋮

リレーショナル・データベースでは，ＥＭＰは関係名となり，ＪＯＨＮ，２５，＄５００はそれぞれ従業員名，年齢，給与となる。機械翻訳システムにおいては，辞書か事実に対応すると考えられ，例えばＰＲＯＬＯＧ上で表現すると次のようになるであろう。

Ｅ－Ｊ（ abandon , ut , 捨てる）.

Ｅ－Ｊ（ abstract , a , 抽象的な）.

Ｅ－Ｊ（ accept , ut , 受取る）.

このように，事実とは世の中の事柄を表現したものであり，ＥＤＰ処理において認識されてきたデータベース内のデータそのものであるといえる。

(3) ＰＲＯＬＯＧの大規模な事実への対応

ＰＲＯＬＯＧの２つの記述能力を従来のプログラミング言語，例えばＣＯＢＯＬに対応させて考えると，規則の部分は手続き部（ＰＲＯＣＥＤＵＲＥ　ＤＩＶＩＳＩＯＮ），事実の部分はデータ記述部（ＤＡＴＡ　ＤＩＶＩＳＩＯＮ）に対応する。一般的に事実の部分は，規則の部分に比して大規模となるものである。従って，従来のプログラミング言語では，明示的に外部記憶装置に事実部分の格納を行い，手続部でもその扱いを意識した。従来のプログラミング言語におけるファイル記述，データベース記述は大規模事実への対応手段であるといえる。一方，ＰＲＯＬＯＧにおいては，事実は，規則とともにあくまで論理的な意味記述であるため，それが実際にコンピュータ上にどのように格納され，処理されるかは利用者の手の届く範囲外にある。一般的なＰＲＯＬＯＧ言語の開発においては，事実と規則は特に区別されることなく，ともに主記憶上に展開されて処理される。実際的には，主記憶の実記憶容量には限りがあるため，仮想記憶空間上に持たれることとなる。もし，大規模な事実を探索しながら処理が進められるならば主記憶の広い範囲での参照がなされる。その場合，実記憶空間には事

実の一部しか持てないので，仮想記憶空間の待避記憶装置と実記憶との間に頻繁な入出力動作がひき起される。このことは，大規模な事実に対するＰＲＯＬＯＧ言語の適応性が悪いことを示している。

一方，最近それらの点に着目してＰＲＯＬＯＧ言語の実現方法の改良がなされている。

その１つに，従来技術であるデータベースとＰＲＯＬＯＧの結合があげられる。この試みは，事実部分の全て，または大半をデータベース上に格納しようというものである。そしてデータベースの持つ強力なデータ検索機能を利用して，効率良く目的データを探索しようというものである。次にその一例を示す。

```
GO(*FROM,*TO):-
      DIA(*FROM,*TO,*TDPT,*TARR),
      PRINT(*FROM,*TO,*TDPT,*TARR).
GO(*FROM,*TO):-
      DIA(*FROM,*STOP,*TDPT,*TARRSTOP),
      DIA(*STOP,*TO,*TDPTSTOP,*TARR),
      <(*TARRSTOP,*TDPTSTOP),PRINT(*FROM,*STOP,*TDPT,*TARRSTOP),
      DISP("  そこで乗り換え  "),
      PRINT(*STOP,*TO,*TDPTSTOP,*TARR).
GO(*FROM,*TO):-
      DIA(*FROM,*STOP1,*TDPT,*TARRSTOP1),
      DIA(*STOP1,*STOP2,*TDPTSTOP1,*TDPTSTOP2),
      DIA(*STOP2,*TO,*TDPTSTOP2,*TARR),
      <(*TARRSTOP1,*TDPTSTOP1),<(*TARRSTOP2,*TDPTSTOP2),
      PRINT(*FROM,*STOP1,*TDPT,*TARRSTOP1),
      DISP("  そこで一度乗り換え，  "),
      PRINT(*STOP1,*STOP2,*TDPTSTOP1,*TDPTSTOP2),
      DISP("  もう一度乗り換え，  "),
      PRINT(*STOP2,*TO,*TDPTSTOP2,*TARR).
PRINT(*FROM,*TO,*TDPT,*TARR):-
      DISPLAY(*FROM,"を",*TDPT,"に出発し",*TO,
            "に",*TARR,"に着きます。").
DIA(東京,大阪,０６００,０９００).   ┐
DIA(東京,大阪,０８００,１１００).   │
DIA(東京,大阪,１０００,１３００).   │
DIA(東京,大阪,１２００,１５００).   │
DIA(大阪,広島,０８００,１０００).   │
DIA(大阪,広島,１０００,１２００).   │
DIA(大阪,広島,１２００,１４００).   ├ assertion (fact)
DIA(大阪,広島,１４００,１６００).   │
DIA(広島,福岡,０９００,１１００).   │
DIA(広島,福岡,１１００,１３００).   │
DIA(広島,福岡,１３００,１５００).   │
DIA(広島,福岡,１５００,１７００).   │
DIA(広島,福岡,１７００,１９００).   ┘
```

図3.5－1　時刻表を検索するプログラム

```
> :-GO(東京,広島),FAIL.
東京を １０００ に出発し 大阪 に １３００ に着きます.
  そこで乗り換え  大阪 を １４００ に出発し 広島 に １６００ に着きます.
東京を ０８００ に出発し 大阪 に １１００ に着きます.
  そこで乗り換え  大阪 を １４００ に出発し 広島 に １６００ に着きます.
東京を ０８００ に出発し 大阪 に １１００ に着きます.
  そこで乗り換え  大阪 を １２００ に出発し 広島 に １４００ に着きます.
東京を ０６００ に出発し 大阪 に ０９００ に着きます.
  そこで乗り換え  大阪 を １４００ に出発し 広島 に １６００ に着きます.
東京を ０６００ に出発し 大阪 に ０９００ に着きます.
  そこで乗り換え  大阪 を １２００ に出発し 広島 に １４００ に着きます.
NO!

>
```

図 3.5－2　時刻表問い合せ実行例

```
:-INQ_OPEN("/PROLOG/DATABAS2").  ①
GO(*FROM,*TO):-
      DIA(*FROM,*TO,*TDPT,*TARR),
      PRINT(*FROM,*TO,*TDPT,*TARR).
GO(*FROM,*TO):-
      DIA(*FROM,*STOP,*TDPT,*TARRSTOP),
      DIA(*STOP,*TO,*TDPTSTOP,*TARR),
      <(*TARRSTOP,*TDPTSTOP),PRINT(*FROM,*STOP,*TDPT,*TARRSTOP),
      DISP("  そこで乗り換え  "),
      PRINT(*STOP,*TO,*TDPTSTOP,*TARR).
GO(*FROM,*TO):-
      DIA(*FROM,*STOP1,*TDPT,*TARRSTOP1),
      DIA(*STOP1,*STOP2,*TDPTSTOP1,*TDPTSTOP2),
      DIA(*STOP2,*TO,*TDPTSTOP2,*TARR),
      <(*TARRSTOP1,*TDPTSTOP1),<(*TARRSTOP2,*TDPTSTOP2),
      PRINT(*FROM,*STOP1,*TDPT,*TARRSTOP1),
      DISP("  そこで一度乗り換え,  "),
      PRINT(*STOP1,*STOP2,*TDPTSTOP1,*TDPTSTOP2),
      DISP("  もう一度乗り換え,  "),
      PRINT(*STOP2,*TO,*TDPTSTOP2,*TARR).
PRINT(*FROM,*TO,*TDPT,*TARR):-
      DISPLAY(*FROM,"を ",*TDPT,"に出発し ",*TO,
            "に ",*TARR,"に着きます。 ").
DIA(*A,*B,*C,*D):-CALL_INQ(DIA,*A,*B,*C,*D).  ②
```

図 3.5－3　データベースを組込んだプログラム例

PREDICATE NAME DIA

ARGUMENT	
01	東京
02	大阪
03	1200
04	1500

PREDICATE NAME DIA

ARGUMENT	
01	大阪
02	広島
03	1400
04	1600

PREDICATE NAME DIA

ARGUMENT	
01	広島
02	福岡
03	1300
04	1500

PREDICATE NAME DIA

ARGUMENT	
01	広島
02	福岡
03	1500
04	1700

・
・
・

図 3.5－4　データベース内の事実の内容

図3.5－1は，東京から福岡までの時刻表に対して，出発地と到着地を入力してどのように電車を乗り換えていけば良いかを知る簡単なプログラムである。また，図3.5－2は出発地に東京，到着地に広島を指定してどのような電車のダイアがあるかを問合せている例である。もし，日本全国の列車のダイアのすべてに対する問合せを行いたいとするならば，図3.5－1におけるＤＩＡの定義に全国の列車を記述しなければならない。そこで，その部分に全てをデータベースに持った例を図3.5－3，4に示す。図中①は，使用するデータベースのＯＰＥＮを行う指定であり，②は，ＤＩＡの事実の定義がすべてデータベース上にあり，それから呼出してデータを利用することを示している。

(4) 事実部分の探索時間

ＰＲＯＬＯＧの実行時間は，大規模な事実部分の探索に要する時間に大きく左右される。

今，ここで事実部分を主記憶に展開して持った場合と，データベース上に持った場合との探索時間の比較を行ってみる。

(a) 主記憶に持つ場合

主記憶に大規模な事実を持ち，その大部分は二次記憶上のプログラム待避エリア上に格納されるものと仮定する。また，その待避エリアとのやりとりに行われる単位であるページに，ｎ個の事実定義が格納され，全体の事実定義の個数をＮとするとＩＯ回数は，

$$\begin{cases} Sio \leq \dfrac{N}{n} \\ Sio\ (\text{平均}) = \dfrac{N}{2n} \end{cases} \quad (\text{Sioは，ディスクIO回数})$$

となる。

今，Ｎ＝１４万，ｎ＝１０とすると，

$$Sio = \frac{14 \times 10^4}{2 \times 10} = 7000\text{回}$$

１ディスクＩＯ時間を３５m secとすると，事実に対する１探索時間は，

$$7000 \times 0.035(s) = 245(s)$$

となり，事実部分のすべてが待避エリア上に持たれるとすると，1回ごとの事実部分のマッチングに２４５秒を要する。

(b) データベース上に持つ場合

事実部分のすべてをデータベース上にインデックス付きで持つと仮定する。この場合，ディスクＩＯ回数は，インデックスの深さと等しくなる。（図3.5－5参照）

ＩＯの単位であるデータベース・ページにｎ個の事実定義が格納され，全体の事実定義の個数をＮとすると，ＩＯ回数は，

$$DBio = \log_n N$$ （DBioはディスクＩＯ回数）

となる。

今，Ｎ＝１４万，ｎ＝１０とすると，

$$DBio = \log_{10} 14 \times 10^4 \fallingdotseq 6回$$

1ディスクＩＯ時間を３５m secとすると事実に対する1探索時間は，

$$6 \times 0.035_{(s)} = 0.21_{(s)}$$

となり，仮想空間を利用して主記憶に事実を持つよりも，データベース上に事実を持つ方が効率上有利であるといえる。

3.5.2 ＰＲＯＬＯＧとデータベースによる機械翻訳の実験

(1) 実験の概要

本実験はＰＲＯＬＯＧ言語からデータベース機能を利用できる点に着目し，現在利用可能となっている英日辞書データのすべてをデータベースに格納し，それを利用した機械翻訳を行うことを目的とする。

(a) 実験のステップ

実験の処理ステップは次の通りである。

① 辞書データを，データベースに格納できる形式に編集する。

ＣＯＢＯＬでプログラムを組み実行。約７００ステップ・コーデ

ィング。

② データベース格納ユーティリティを使用してデータベースに格納。所要時間は約２０分。データベース格納後のデータ量は，約３０ＭＢ。データ圧縮率は，約３５％。

③ ＴＳＳ環境よりＰＲＯＬＯＧを利用して翻訳実験を行う。

それらの処理フロー図を図3.5－6に示す。

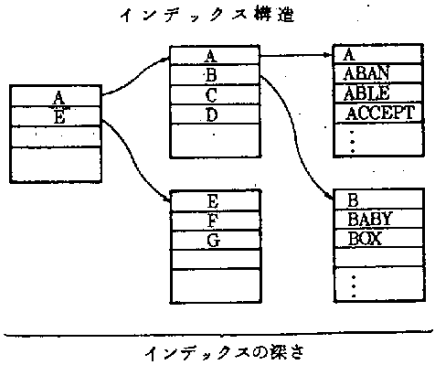


図3.5－5　インデックス構造

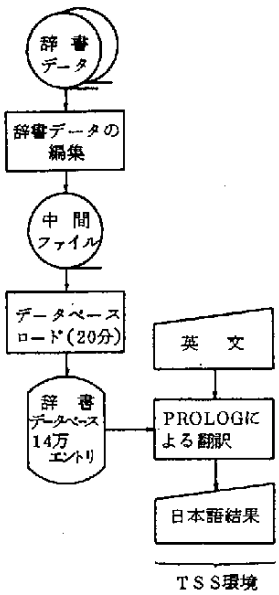


図3.5－6　実験のステップ・フロー

(b) 実験環境

本実験環境は，次の通りである。

ハードウェア
- システム名　ACOSシステム650
- 主記憶メモリ　4MB
- ディスク装置　317MB×2　200MB×1

ソフトウェア
- データベース　INQ(INformation Query)[5]
- PROLOG　Shape Up[6]

(2) 辞書データベースの作成

辞書データベースに対して述語名をDICTとし，次の構造とした。

DICT(見出し語，品詞，日本語訳，補足説明，分野)

以下その作成内容を示す。

(a) 辞書レコードの作成

辞書レコードは，PROLOGからの処理の容易性を考慮し，リスト構造を含まない単純データ形式とした。また辞書語義中の説明番号ごとに，最大2個までをデータベースに登録し，3個以降は切り捨てた。その関連を図3.5－7に示す。

(b) 見出し語

見出し語は，中間ピリオドのない英大文字とした。辞書にはカンマで2つの見出し文字を意味する場合があり，2つの見出し語のレコードを作る必要がある。しかし，今回はそのままとした。

例　Zaffer, Zaffre

⇩

ZAFFER

ZAFFRE

＊＊ fear〔 fir／fiə 〕n. 1.恐れ，恐怖．2.心配，懸念，気遣い，心配の種‖不安，危険性〔悪いことの起る〕。
3.畏(い)怖，畏敬〔神などに対する〕。

⇩ 抜き出し

DICT(FEAR, N, 恐れ,,)
DICT(FEAR, N, 恐怖,,)
DICT(FEAR, N, 心配,,)
DICT(FEAR, N, 懸念,,)
DICT(FEAR, N, 畏怖,,)
DICT(FEAR, N, 畏敬, 神などに対する,)
補足説明

図3.5－7　辞書レコードの抜き出し

(c) 品　　詞

品詞コードは，ＰＲＯＬＯＧで記述した構文解析プログラムと密接な関係があり，互いにインターフェイスを取りながら表3.5－1のように決定した。

(d) 日本語訳

日本語訳については，助詞を含まない原形の形で格納した。

(e) 補足説明

辞書中には，意味的な説明をするために，補足説明が付されている。例えば，draw では〔人の注意，興味などを〕引く，〔剣を〕抜く，〔水を〕汲む，〔利益を〕もたらす，〔線を〕引く，〔手形

を〕振出す，……等がある。

これらは，将来意味処理をする上で非常に有益であると考えられ，そのカッコ内のデータを取り出した。おそらく意味シソーラス述語を利用したその取扱いが予想される。

(f) 分　　野

辞書の中には分野によっては訳が異なるものが少なくない。例えばoperationは，〖医〗では「手術」と訳し，〖軍〗では「作戦」と訳す。分野が異なればそれらは全く別の意味をもつため，分野に対する考慮が将来必要と思われ，その部分のデータを取り出した。

表3.5－1　品詞コード表

品詞コード	意　味	品詞コード	意　味
A	形容詞	EPP	過去分詞
AD	副　詞	PREF	接頭辞
X	助動詞	PREP	前置詞
CONJ	接続詞	PRON	氏名詞
DEF	定冠詞	REL	関係代名詞
INDEF	不定冠詞	SUF	接尾辞
INT	間投詞	VT	他動詞
N	名　詞	VI	自動詞

(3) ＰＲＯＬＯＧによる翻訳プログラム

本実験で使用したＰＲＯＬＯＧプログラムとデータベースの構成を図3.5－8に示す。

プログラムを作成するに当って文献7)及び8)を参考にした。

本実験システムでは，トランスファ方式を採用している。まず入力された英文を構文解析し，中間言語形式として英文の句構造構文木を

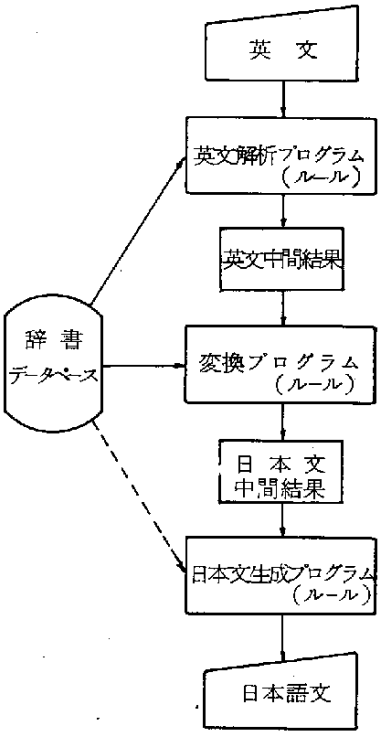


図 3.5－8　PROLOGプログラムとデータベースの構成

出力する。変換プログラムは，それを目的言語の句構造構文木に変換し，生成プログラムは，それを最終的な目的言語である日本文にする。

(a)　英文構文解析プログラム

ＰＲＯＬＯＧで文法規則を記述するのに適した文法規則に文脈自由文法（ＣＦＧ；Context Free Grammer ）がある。ＣＦＧでは，文法の個々の部分表現をしておき，それらが集まってより広範な書き換え規則となり得る。例えば，

Ｓ→ＳＳ・ＥＮＤ

（文はセンテンスとピリオドから成る）

ＳＳ→ＳＵＢ・ＰＲＥＤ

（センテンスは，主部と述部から成る）

ＳＵＢ→ＮＰ

（主部は，名詞節より成る）

NP→DET・N

（名詞節は，不定冠詞と名詞から成る）

NP→N

（あるいは，名詞節は名詞から成ってもよい）

PRED→VP・DOB

（述部は，動詞節と直接目的語から成る）

VP→V

（動詞節は，動詞から成る）

DOB→NP

（直接目的語は，名詞節から成る）

N→We，　Computer

（名詞には，“We”や“Computer”がある）

DET→a

（不定冠詞には“a”がある）

V→buy

（動詞には“buy”がある）

END→．

（ピリオド）

が与えられると，〔We buy a computer〕は図3.5－9の構造に分解される。

PROLOGにおいても，このCFGと同一の形式で表現することが可能であり，ルールそのものがプログラムとなる。ただ，CFGと異なる点は，PROLOGが変数を含みそれを引数として各種の関数による広範な処理機能を備えることにある。通常CFGをベースにした文法規則ではその記述能力が不足するため書き換え規則に手続きが付加されるが，それらはPROLOGにおける変数と，

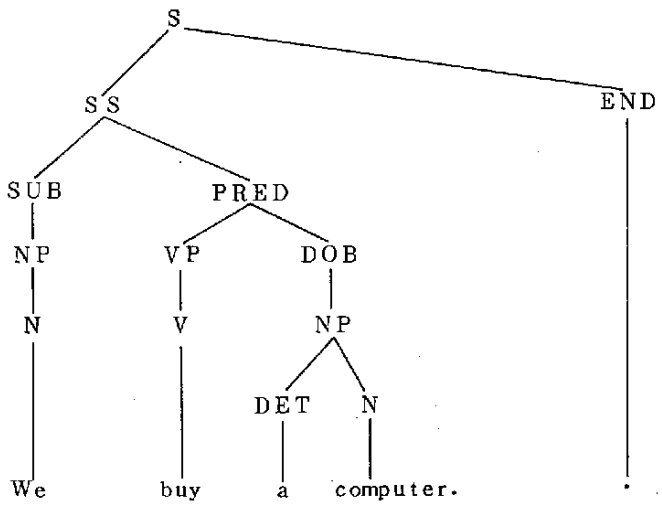


図3.5－9　構文解析木の例

それをもとにした関数処理によって代替できる。次の本実験システムにおけるルールの一部分による英文解析例を示す。

①は，辞書データベースのＯＰＥＮを行う指定である。②は，ＲＥＡＤＩＮ述語により端末より英文を読み込み，それをリスト構造に直して＊Ａに代入し，ＥＳに渡すことを示している。ＥＳでは，代入された＊Ａの内容を下のラインに続く各ルールを用いて解析し，その解析結果をＤＩＳＰＬＡＹ述語に渡す。ＤＩＳＰＬＡＹ述語でその内容を表示し，同時にＣＯＮＶ－ＧＯという変換プログラムに解析結果を渡す。③は，文が主部と述部から成ることを示している。④は，主部が名詞節から成ることを示している。⑤は述部を示し，動詞と目的語から成る事を示している。⑥は，名詞節の定義を行っている。名詞節がＥＮＰ１からＥＮＰ３まであるのは，名詞が複数個並んだ場合の処置であり，ＰＲＯＬＯＧにおけるレフト・リカーシブを避けるための考慮である。⑦は，辞書内のデータでその品詞がＩＮＤＥＦであれば不定冠詞であることを示している。⑧は，同

```
①…… :-INQ_OPEN('      /JISHO/DB/DICT').
     /*      PARSING PROGRAM      */
②……GO:-READIN(*A),!,ES(*A,*B,*C),DISPLAY(*C),!,CONV_GO(*C).
    CHECK([]).
③…{ES(*A,*B,[ES,*P]):-ESS(*A,*B,*P).
   {ESS(*A,*B,[ESS,*P,*Q]):-ESUB(*A,*C,*P),EDES(*C,*B,*Q).
④…{ESUB(*A,*B,[ESUB,*P]):-ESUB1(*A,*B,*P).
   {ESUB1(*A,*B,[ESUB1,*P]):-ENP(*A,*B,*P).
   {EDES(*A,*B,[EDES,*P]):-EPRED(*A,*B,*P),CHECK(*B).
   {EPRED(*A,*B,[EPRED,*P]):-EPREDX(*A,*B,*P).
⑤…{EPREDX(*A,*B,*C):-EPRED1(*A,*B,*C).
   {EPRED1(*A,*B,[EPRED1,*P,*Q]):-EV(*A,*C,*P),EDOB(*C,*B,*Q).
   {EDOB(*A,*B,[EDOB,*P]):-ENP(*A,*B,*P).
   {EV(*A,*B,[EV,*P]):-EVT(*A,*B,*P).
   {ENP(*A,*B,[ENP,*P]):-ENP1(*A,*B,*P).
   {ENP1(*A,*B,[ENP1,*P,*Q]):-EDET(*A,*C,*P),ENP2(*C,*B,*Q).
⑥…{ENP1(*A,*B,[ENP1,*P]):-ENP2(*A,*B,*P).
   {ENP2(*A,*B,[ENP2,*P]):-ENP3(*A,*B,*P).
   {ENP3(*A,*B,[ENP3,*P]):-EN(*A,*B,*P).
⑦……EDET(*A,*B,[EDET,*P]):-DICT(*A,INDEF,*B,*P).
⑧……EN(*A,*B,[EN,*P]):-DICT(*A,N,*B,*P).
⑨……EVT(*A,*B,[EVT,*P]):-DICT(*A,VT,*B,*P).
⑩……DICT([*A$*B],*C,*B,*A):-!,CALL_INQ(DICT,*A,*Z,*D,*X,*Y),==(*Z,*C).
```

(a)　英文解析プログラム

```
⑪……> :-GO.
⑫……? WE BUY A COMPUTER.
⑬…{[WE,BUY,A,COMPUTER]
   {[ES,[ESS,[ESUB,[ESUB1,[ENP,[ENP1,[ENP2,[ENP3,[EN,WE]]]]]],[EDES,[EPRED,[EPRED1
   {,[EV,[EVT,BUY]],[EDOB,[ENP,[ENP1,[EDET,A],[ENP2,[ENP3,[EN,COMPUTER]]]]]]]]]]]]]
```

(b)　構文解析実行例

図 3.5－10　　英 文 解 析 例

じく辞書内の品詞がNであれば名詞であることを示している。⑨は，動詞であることを示している。⑩は，辞書引きのラインであり，〔＊A$＊B〕は，リストの第1データを＊Aに代入し，残りのすべてを＊Bに代入する。⑪は，述語名GOのラインを実行する。⑫はREADIN述語からの入力促進に対して英文の入力を行ったものである。⑬は，入力した英文に対する解析結果であり，分り易く図示すると図3.5－11のようになる。

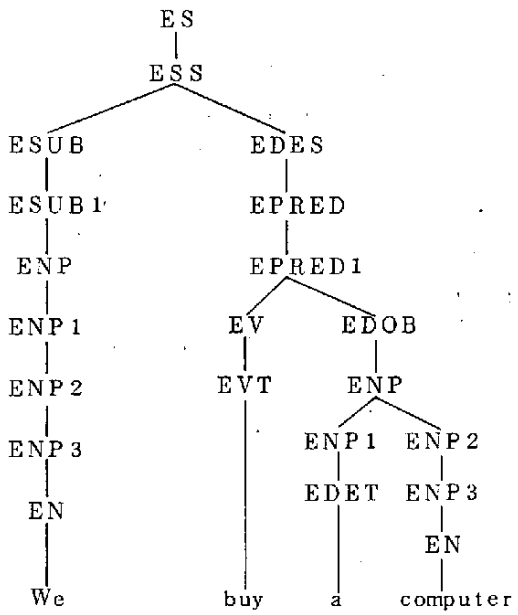


図3.5－11　英文解析木

(b)　変換プログラム

トランスファ方式の翻訳システムでは，ソース言語の中間表現形式を目的言語の中間表現形式へ変換する。本実験システムでは，前述のように中間表現として句構造構文木を採用している。従って変換プログラムは，木から木への変換プログラムとなる。この処理においてもPROLOGの持つパターンマッチング機能を有効に利用できる。変換プログラムは，変換規則に従って，英文の構文木を上から順次調べていき，対応する日本文の構文木を上から下へ段階的

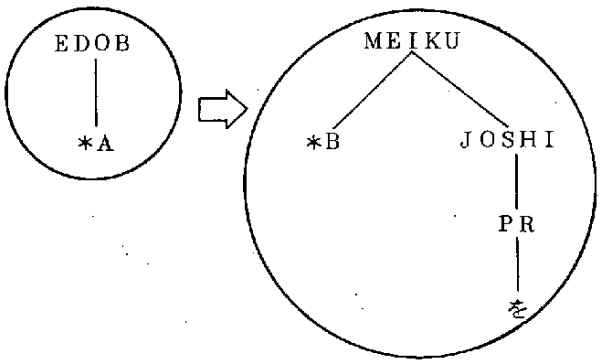


図3.5－12　木の変換例

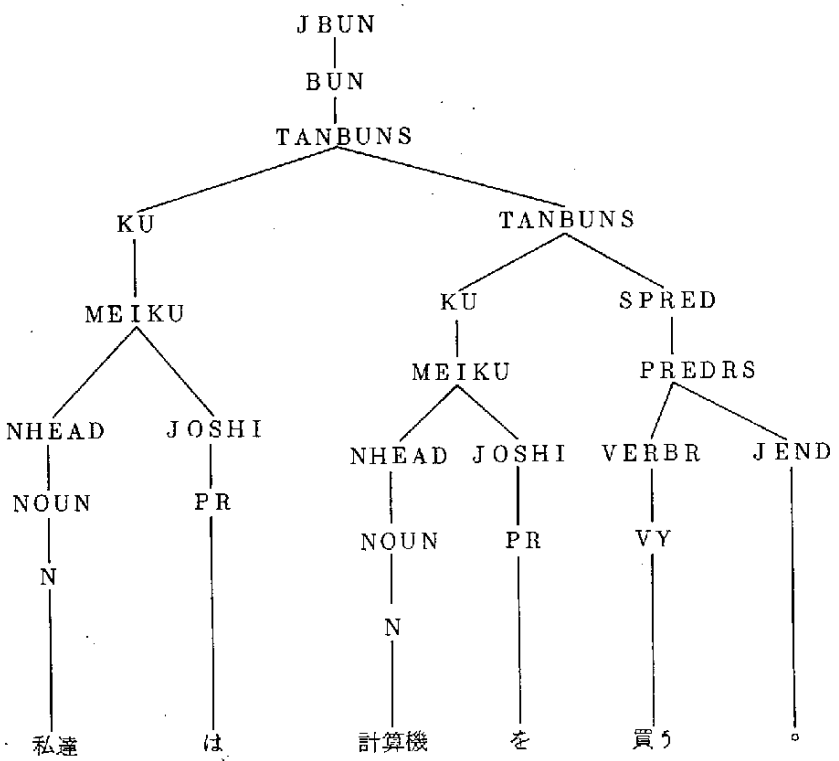


図3.5－14　変換結果木

に組立てていくという変換過程をとる。例えば図3.5－12のようになる。つまりＴＫＵという述語の第１引数が〔ＥＤＯＢ，＊Ａ〕であるならば，変数＊Ａで示される部分構造をＴＥＭＰで示される変換規則によって変換し，その結界を＊Ｂに代入する。図3.5－１３に〔We buy a computer〕の英文中間表現に対する日本語中間表現を作る

```
/*      CONVERSION PROGRAM      */
TVT([EVT,*A],[VY,*B]):-DICT2(*A,VT,*B).
TENN([EN,*A],[N,*B]):-DICT2(*A,N,*B).
/*      ##      */
CONV_GO(*Y):-TRA(*Y,*Z),DISPLAY(*Z),GENERATE(*Z).
TRA([ES,*A],[JBUN,*B]):-TBE(*A,*B).
TBE([ESS,*A,*B],[BUN,*C]):-TBUN(*A,*B,*C).
TBUN([ESUB,*A],[EDES,[EPRED,*B]],[TANBUNS,[KU,*C],*D]):-TKU(*A,*C)
,TCASE(*B,*D).
TCASE([EPRED1,*A,*B],[TANBUNS,[KU,*C],*D]):-TKU(*B,*C),TSPRED(*A,*D).
TSPRED([EV,*A],[SPRED,[PREDRS,[VERBR,[VY$*B]],[JEND,,  ]]]):-TVT(*A,[VY$*B]).
TKU([ESUB1,*A],[MEIKU,*B,[JOSHI,[PR,は ]]]):-TENP(*A,*B).
TKU([EDOB,*A],[MEIKU,*B,[JOSHI,[PR,を ]]]):-TENP(*A,*B).
TENP([ENP,[ENP1,[EDET,*A],*B]],[NHEAD,*C]):-TENP2(*B,*C).
TENP([ENP,[ENP1,*A]],[NHEAD,*B]):-TENP2(*A,*B).
TENP2([ENP2,[ENP3,*A]],[NOUN,*D]):-TENN(*A,*D),(*A).
DICT2(*A,*B,*C):-CALL_INQ(DICT,*A,*Z,*C,*D,*E),==(*Z,*B),!.
```

```
[JBUN,[BUN,[TANBUNS,[KU,[MEIKU,[NHEAD,[NOUN,[N,私達 ]]],[JOSHI,[PR,は ]]]],[TANBU
NS,[KU,[MEIKU,[NHEAD,[NOUN,[N,計算機 ]]],[JOSHI,[PR,を ]]]],[SPRED,[PREDRS,[VERBR
,[VY,買う ]],[JEND,,  ]]]]]]]
```

図3.5－13　変換プログラムと変換結果

プログラム例とその結果を示す。

(c) 日本文生成プログラム

日本文生成プログラムは，英文解析プログラムと全く同じ考え方で作られている。つまり，英文解析プログラムは一連の英文を読み込み，構文解析木を作ることであった。ＰＲＯＬＯＧのもう一つの特長として可逆性がある。この可逆性を応用して構文解析木を与えてその出力として一連の文が得られないだろうかと考えた。この考えをもとにして，日本文生成プログラムを作った。つまり，日本文生成プログラムは，入力された一連の日本文を解析して構文解析木を作る事を考えて作られている。そして，完成後に，変換プログラムと結合され，結果的に日本文生成プログラムとなった。このことは，本実験システムが，ＰＲＯＬＯＧの可逆性の応用により，日英翻訳システムとしても基本的には利用できることを示している。図3.5−15に日本文生成プログラムと生成結果を示す。

```
/*      GENERATION PROGRAM      */
JEND([*P,NIL],NIL,[JEND,*P]).
PR([*P$*A],*A,[PR,*P]).
N([*P$*A],*A,[N,*P]).
VR([*P$*A],*A,[VY,*P]).
NBUN([, $*B]).
NBUN([*A$*B]):-KAKKO(*A,*B).
KAKKO(*A,*B):-DISP(*A),NBUN(*B).
JOSHI(*A,*B,[JOSHI,*P]):-PR(*A,*B,*P).
VERBR(*A,*B,[VERBR,*P]):-VR(*A,*B,*P).
PREDRS(*A,*B,[PREDRS,*P,*Q]):-VERBR(*A,*C,*P),JEND(*C,*B,*Q).
NOUN(*A,*B,[NOUN,*P]):-N(*A,*B,*P).
NHEAD(*A,*B,[NHEAD,*P]):-NOUN(*A,*B,*P).
MEIKU(*A,*B,[MEIKU,*P,*Q]):-NHEAD(*A,*C,*P),XXX4(*C,*B,*Q).
XXX4(*A,*B,*C):-JOSHI(*A,*B,*C).
KU(*A,*B,[KU,*P]):-MEIKU(*A,*B,*P).
TANBUNS(*A,*B,[TANBUNS,*P,*Q]):-KU(*A,*C,*P),XXX1(*C,*B,*Q).
XXX1(*A,*B,*C):-TANBUNS(*A,*B,*C).
XXX1(*A,*B,*C):-SPRED(*A,*B,*C).
SPRED(*A,*B,[SPRED,*P]):-PREDRS(*A,*B,*P).
BUN(*A,*B,[BUN,*P]):-TANBUNS(*A,*B,*P).
JBUN(*A,*B,[JBUN,*P]):-BUN(*A,*B,*P).
/*      ##      */
GENERATE(*X):-JBUN(*Z,*Y,*X),NBUN(*Z).
```

(a) 日本文生成プログラム

```
私達は計算機を習う
>
```

(b) 日本文生成結果

図3.5−15 日本文生成プログラムとその結果

3.5.3 実験結果

本実験は，１４万エントリの辞書データベースを持つことを特長としており，その特長を活かして簡易な構文ではあるが，難解な単語を選んで翻訳実験を行った。

実験には次の文を使用した。

① We buy a computer.

② The conference denies the sovereignty of Northumbria.

③ The institution gives the doctor a electroencephalograph.

④ The impressionism directs the trend of art.

⑤ The poverty cause the revolution which destroy the existence.

⑥ The bureaucrat controls the agency which collects the information.

⑦ With the advancement, the computer gives the society convenience.

以下その結果を示す。

```
> :-GO.
? WE BUY A COMPUTER.
[WE.BUY.A.COMPUTER]
**********  THE ENGLISH PARSING TREE **********
[ES.[ESS.[ESUB.[ESUB1.[ENP.[ENP1.[ENP2.[ENP3.[EN.WE]]]]]]].[EDES.[EPRED.[EPRED1
.[EV.[EVT.BUY]].[EDOB.[ENP.[ENP1.[EDET.A].[ENP2.[ENP3.[EN.COMPUTER]]]]]]]]]]]]
********** THE JAPANESE PARSING TREE **********
[JBUN.[BUN.[TANBUNS.[KU.[MEIKU.[NHEAD.[NOUN.[N.私達 ]]].[JOSHI.[PR.は ]]]].[TANBU
NS.[KU.[MEIKU.[NHEAD.[NOUN.[N.計算機 ]]].[JOSHI.[PR.を ]]]].[SPRED.[PREDRS.[VERBR
.[VY.買う ]].[JEND.. ]]]]]]]
********** THE GENERATED JAPANESE SENTENCE **********
私達は計算機を買う
>
```

図3.5－16　①の結果


```
> :-GO.
? THE CONFERENCE DENY THE SOVEREIGNTY OF NORTHUMBRIA.
[THE.CONFERENCE.DENY.THE.SOVEREIGNTY.OF.NORTHUMBRIA]
**********  THE ENGLISH PARSING TREE **********
[ES.[ESS.[ESUB.[ESUB1.[ENP.[ENP1.[EDET.THE].[ENP2.[ENP3.[EN.CONFERENCE]]]]]].[
EDES.[EPRED.[EPRED1.[EV.[EVT.DENY]].[EDOB.[ENP.[ENP1.[EDET.THE].[ENP2.[ENP3.[EN
.SOVEREIGNTY]]]].[EPOSTMOD.[EOFP.[EOF.OF].[ENP.[ENP1.[ENP2.[ENP3.[EN.NORTHUMBRI
A]]]]]]]]]]]]]]]]]
********** THE JAPANESE PARSING TREE **********
[JBUN.[BUN.[TANBUNS.[KU.[MEIKU.[NHEAD.[NOUN.[N.協議会 ]]].[JOSHI.[PR.は ]]]].[TAN
BUNS.[KU.[MEIKU.[NHEAD.[RENKU.[NHEAD.[NOUN.[N.ノーサンブリア ]]].[RJOSHI.[NO.の ]
]].[NHEAD.[NOUN.[N.主権 ]]]].[JOSHI.[PR.を ]]]].[SPRED.[PREDRS.[VERBR.[VY.否定す
る ]].[JEND.. ]]]]]]]
********** THE GENERATED JAPANESE SENTENCE **********
協議会はノーサンブリアの主権を否定する
>
```

図3.5－17　②の結果

```
> :-GO.
? THE INSTITUTION GIVE THE DOCTOR THE ELECTROENCEPHALOGRAPH.
[THE,INSTITUTION,GIVE,THE,DOCTOR,THE,ELECTROENCEPHALOGRAPH]
**********  THE ENGLISH PARSING TREE **********
[ES,[ESS,[ESUB,[ESUB1,[ENP,[ENP1,[EDET,THE],[ENP2,[ENP3,[EN,INSTITUTION]]]]]],
[EDES,[EPRED,[EPRED1,[EV,[EVT,GIVE]],[EIOB,[ENP,[ENP1,[EDET,THE],[ENP2,[ENP3,[E
N,DOCTOR]]]]]],[EDOB,[ENP,[ENP1,[EDET,THE],[ENP2,[ENP3,[EN,ELECTROENCEPHALOGRAP
H]]]]]]]]]]
********** THE JAPANESE PARSING TREE **********
[JBUN,[BUN,[TANBUNS,[KU,[MEIKU,[NHEAD,[NOUN,[N,公共団体]]],[JOSHI,[PR,は]]]],[T
ANBUNS,[KU,[MEIKU,[NHEAD,[NOUN,[N,医者]]],[JOSHI,[PR,に]]]],[TANBUNS,[KU,[MEIKU
,[NHEAD,[NOUN,[N,脳波記録装置]]],[JOSHI,[PR,を]]]],[SPRED,[PREDRS,[VERBR,[VY,与
える]],[JEND,。]]]]]]]]
********** THE GENERATED JAPANESE SENTENCE **********
公共団体は医者に脳波記録装置を与える
>
```

図3.5－18　③の結果


```
> :-GO.
? THE IMPRESSIONISM DIRECT THE TREND OF ART.
[THE,IMPRESSIONISM,DIRECT,THE,TREND,OF,ART]
**********  THE ENGLISH PARSING TREE **********
[ES,[ESS,[ESUB,[ESUB1,[ENP,[ENP1,[EDET,THE],[ENP2,[ENP3,[EN,IMPRESSIONISM]]]]]]
],[EDES,[EPRED,[EPRED1,[EV,[EVT,DIRECT]],[EDOB,[ENP,[ENP1,[EDET,THE],[ENP2,[ENP
3,[EN,TREND]]]],[EPOSTMOD,[EOFP,[EOF,OF],[ENP,[ENP1,[ENP2,[ENP3,[EN,ART]]]]]]]
]]]]]]
********** THE JAPANESE PARSING TREE **********
[JBUN,[BUN,[TANBUNS,[KU,[MEIKU,[NHEAD,[NOUN,[N,印象主義]]],[JOSHI,[PR,は]]]],[T
ANBUNS;[KU,[MEIKU,[NHEAD,[RENKU,[NHEAD,[NOUN,[N,芸術]]],[RJOSHI,[NO,の]]],[NHEA
D,[NOUN,[N,向き]]]],[JOSHI,[PR,を]]]],[SPRED,[PREDRS,[VERBR,[VY,指導する]],[JEN
D,。]]]]]]]
********** THE GENERATED JAPANESE SENTENCE **********
印象主義は芸術の向きを指導する
>
```

図3.5－19　④の結果

```
> :-GO.
? THE POVERTY CAUSE THE REVOLUTION WHICH DESTROY THE EXISTENCE.
[THE.POVERTY.CAUSE.THE.REVOLUTION.WHICH.DESTROY.THE.EXISTENCE]
**********  THE ENGLISH PARSING TREE **********
[ES.[ESS.[ESUB.[ESUB1.[ENP.[ENP1.[EDET.THE].[ENP2.[ENP3.[EN.POVERTY]]]]]].[EDE
S.[EPRED.[EPRED1.[EV.[EVT.CAUSE]].[EDOB.[ENP.[ENP1.[EDET.THE].[ENP2.[ENP3.[EN.R
EVOLUTION]]]].[EPOSTMOD.[ERELWHICH.[EWHICH.WHICH].[EPRED.[EPRED1.[EV.[EVT.DESTR
OY]].[EDOB.[ENP.[ENP1.[EDET.THE].[ENP2.[ENP3.[EN.EXISTENCE]]]]]]]]]]]]]]]]]]]]]]
********** THE JAPANESE PARSING TREE **********
[JBUN.[BUN.[TANBUNS.[KU.[MEIKU.[NHEAD.[NOUN.[N.貧乏 ]]].[JOSHI.[PR.は ]]]].[TANBU
NS.[KU.[MEIKU.[NHEAD.[EM.[KU.[MEIKU.[NHEAD.[NOUN.[N:実在 ]]].[JOSHI.[PR.を ]]]].[
EMPRED.[PREDRE.[VERBR.[VY.破壊する ]]]]].[NHEAD.[NOUN.[N.革命 ]]]].[JOSHI.[PR.を ]
]]].[SPRED.[PREDRS.[VERBR.[VY.の原因となる ]].[JEND.. ]]]]]]]
********** THE GENERATED JAPANESE SENTENCE **********
貧乏は実在を破壊する革命をの原因となる
>
```

図3.5－20　⑤の結果

```
> :-GO.
? THE BUREAUCRAT CONTROL THE AGENCY WHICH COLLECT THE INFORMATION.
[THE.BUREAUCRAT.CONTROL.THE.AGENCY.WHICH.COLLECT.THE.INFORMATION]
**********  THE ENGLISH PARSING TREE **********
[ES.[ESS.[ESUB.[ESUB1.[ENP.[ENP1.[EDET.THE].[ENP2.[ENP3.[EN.BUREAUCRAT]]]]]].[
EDES.[EPRED.[EPRED1.[EV.[EVT.CONTROL]].[EDOB.[ENP.[ENP1.[EDET.THE].[ENP2.[ENP3.
[EN.AGENCY]]]].[EPOSTMOD.[ERELWHICH.[EWHICH.WHICH].[EPRED.[EPRED1.[EV.[EVT.COLL
ECT]].[EDOB.[ENP.[ENP1.[EDET.THE].[ENP2.[ENP3.[EN.INFORMATION]]]]]]]]]]]]]]]]]]]]]]
********** THE JAPANESE PARSING TREE **********
[JBUN.[BUN.[TANBUNS.[KU.[MEIKU.[NHEAD.[NOUN.[N.官僚式の人 ]]].[JOSHI.[PR.は ]]]].
[TANBUNS.[KU.[MEIKU.[NHEAD.[EM.[KU.[MEIKU.[NHEAD.[NOUN.[N.通知 ]]].[JOSHI.[PR.を
]]]].[EMPRED.[PREDRE.[VERBR.[VY.集める ]]]]].[NHEAD.[NOUN.[N.働き ]]]].[JOSHI.[PR
.を ]]]].[SPRED.[PREDRS.[VERBR.[VY.支配する ]].[JEND.. ]]]]]]]
********** THE GENERATED JAPANESE SENTENCE **********
官僚式の人は通知を集める働きを支配する
>
```

図3.5－21　⑥の結果

```
> :-GO.
? WITH THE ADVANCEMENT,THE COMPUTER GIVE THE SOCIETY CONVENIENCE.
[WITH,THE,ADVANCEMENT,,,THE,COMPUTER,GIVE,THE,SOCIETY,CONVENIENCE]
**********  THE ENGLISH PARSING TREE **********
[ES,[ESS,[EPREPP,[EPREPP1,[EPREP,WITH],[ENP,[ENP1,[EDET,THE],[ENP2,[ENP3,[EN,AD
VANCEMENT]]]]]]],[ECOM,,],[ESUB,[ESUB1,[ENP,[ENP1,[EDET,THE],[ENP2,[ENP3,[EN,CO
MPUTER]]]]]]],[EDES,[EPRED,[EPRED1,[EV,[EVT,GIVE]],[EIOB,[ENP,[ENP1,[EDET,THE],
[ENP2,[ENP3,[EN,SOCIETY]]]]]],[EDOB,[ENP,[ENP1,[ENP2,[ENP3,[EN,CONVENIENCE]]]]
]]]]]]
********** THE JAPANESE PARSING TREE **********
[JBUN,[BUN,[TANBUNS,[KU,[FUKUKU,[NHEAD,[NOUN,[N,前進 ]]],[JOSHIF,[PREP1,とともに
]]]],[TANBUNS,[KU,[MEIKU,[NHEAD,[NOUN,[N,計算機 ]]],[JOSHI,[PR,は ]]]],[TANBUNS,[
KU,[MEIKU,[NHEAD,[NOUN,[N,社会 ]]],[JOSHI,[PR,に ]]]],[TANBUNS,[KU,[MEIKU,[NHEAD,
[NOUN,[N,便利 ]]],[JOSHI,[PR,を ]]]],[SPRED,[PREDRS,[VERBR,[VY,与える ]],[JEND,, ]
]]]]]]]]
********** THE GENERATED JAPANESE SENTENCE **********
前進とともに計算機は社会に便利を与える
>
```

図 3.5－22　⑦の結果

以上，結果を示したが，本実験システムでは，動詞の語尾変化処理を行っていない。従って動詞は原形で入力している。

3.5.4 実験の評価と今後の課題

(1) 評　価

本実験の評価を次の項目について行う。

(a) 開発の容易性

ＰＲＯＬＯＧとデータベース機能を利用した今回の開発コストは，同様のことをＰＬ／１やＣＯＢＯＬ等を使用した場合の開発コストに比して，１桁の差はあるものと思われる。つまり，文法ルール記述そのものがＰＲＯＬＯＧプラグラムであり，いかに処理を行うかの記述は全く不要で，何を行うかの記述で事足りる。またさらに，エンドユーザーであっても，ＰＲＯＬＯＧの文法は単純であるため，短時間の勉強で使えるまでになれる。ルールを変更したい時に，プログラマに言って変えてもらうよりも自分で変更できる点は魅力である。今回作成した翻訳プログラムのＰＲＯＬＯＧ上でのステップ数は次の通りである。

・英文解析プログラム

ステップ数約１２０　ルール数約１００

・変換プログラム

ステップ数約８０　　ルール数約５０

・日本文生成プログラム

ステップ数約１３０　ルール数約１００

(b) 処理効率

ＰＲＯＬＯＧを使用した事によって，結果的にはTop down 方式となった。従って，ルールが早く見つからない場合には，後戻り処理（バックトラック）により指数関数的に処理時間が増大した。例

えば，前項の７つの文に対する翻訳処理で，最も早いものでは２～３秒でその翻訳結果が出力された。その場合を分析すると，ＰＲＯＬＯＧで記述した文法の順序が，入力した英文にマッチしたものであった。７つの例で最も遅いものでは数分を越えるものもあった。これは，英文に文法の順序が合致していないものであった。

一般にＷＦＳＴ（Well Formed Substring Table）を用いた場合n^3に処理時間が押えられるといわれているが，本プログラムの見直しをして（例えば，cut 処理を入れる等して），効率面でそれに近ずける改良が必要であると思われる。ルールが増える場合には，効率の問題が重大問題となることが明らかである。次のデータベースのアクセスであるが，当初データベースに対するユニフィケーションがまずく，１回のデータベース探索に数秒を要した。当初は，必ず見出し語と品詞に値が代入されてしまい，ＤＢＭＳの中で２つの集合のＡＮＤ処理が行われていた。

もし，品詞に名詞Ｎが代入されると，１０万件以上がヒットし，ＡＮＤ処理に負荷がかかる。そこで，その記述を次の様に改良した。

```
:－ＣＡＬＬ＿ＩＮＱ（ＤＩＣＴ，＊Ａ，Ｎ，＊Ｃ，＊Ｄ，
  ＊Ｅ）．
```

⇩

```
:－ＣＡＬＬ＿ＩＮＱ（ＤＩＣＴ，＊Ａ，＊Ｂ，＊Ｃ，＊Ｄ，
  ＊Ｅ），＝＝（＊Ｂ，Ｎ）．
```

この改良は次の事を意味している。つまり，見出し語で選ばれた高々数件のデータに対して，その内容を読んで，品詞が“Ｎ”であるものをのみ真とする。それ故，品詞がＮである１０万件以上の集合と論理積をとる必要はなくなり，一回のデータベースアクセスは数回のディスクＩＯ動作だけで良くなった。

また，ＰＲＯＬＯＧを利用したことにおいて，効率の問題以外に

プログラム・ループの問題がある。これは，ルールが複雑となり，レフト・リカーシブ状態となった時に発生するものであり，それについても今後の検討を必要としている。

(c) 拡張性

前述のようにＰＲＯＬＯＧプラグラムそのものがルールである。従って，ルールの変更が必要となった場合には，容易にその対応が可能である。一方，従来のプログラミング言語で容易に指定可能だったＩＦ／ＴＨＥＮ／ＥＬＳＥ文と同様の指定は逆に困難となった。特に例外処理を木目細かく指定する事はできないため，cut 処理を組合わせて対処することとなる。このことは，ＰＲＯＬＯＧの良さを失うことにもつながるため，さらに検討を要するものと思われる。今後の方向として，ＰＲＯＬＯＧ言語で処理が困難である点に関しては，従来のプログラミング言語で記述しリンクして利用する方法等が考えられる。

(2) 今後の課題

今後の課題として次の点があげられる。

(a) 効率アップ

先に述べたように，効率の向上を計るための検討を行い，それらを取込む事を行う。

(b) ルールの検証と拡充

現在本実験システムは，やっと平易な構文が解析できたにすぎない。今後は，語尾変化処理，イディオム処理[8]等を順次取込んでいく必要があると考えている。また，数多くの英文に対して翻訳実験を行い，その結果をプログラムにフィードバックし，プログラムの充実を計ることが今後必要である。

(c) 意味処理への対応

現在のプログラムでは全く意味処理を行っていない。例えば実験

結果⑤のＣＡＵＳＥに対する訳語が「の原因となる」となっているが，たまたまこれが辞書の最初に格納されているため，このように訳された。その第２訳語は「引起こす」であった。もし⑤の結果にこれを当てはめれば，その結果は正しいものとなる。この様に意味処理は，今後困難ではあるが段階的に解決されねばならない重要な問題であるといえる。すでに辞書データベースには，今後の意味処理に備えて説明文や分野データを含んでいる。

(d) 辞書データベースの充実

翻訳の良し悪しは辞書データベースの内容で決まるといっても過言ではない。本実験を通じて，より良い機械翻訳のために，どのような情報を辞書データベースに取込む事が必要かということを明らかにする必要がある。

辞書データの整備には時間と労力を多く必要とするため，早期それらが明らかになればなるほど良い翻訳の実現日が近付くこととなる。また同時に，分野毎の専門用語辞書や，現在の一般辞書の拡充がさらに望まれる。それに従って，辞書データベース更新のための勝れたマン・マシン・インターフェイスの実現が期待される。

3.5.6 リスト例

(1) 辞書データベースの出力例

次に辞書データベースに対して，エンドユーザ言語で検索した結果の一部を示す。

＜データの見方＞

```
-----レコード番号 ----------------------------
P－NAME      DICT
  VAL        DICT    01  見出し語
  VAL        DICT    02  品　詞
```

```
SYSTEM ? INQ
   INQ SQL/JIPS VERSION 9.1-00    16:13'49"     031684
OPTION FILE ? /JISHO/DB/OPT

  ------------------------------------------------------------
  : FILE NO : FILE NAME     : RECORD CNT : DATABASE NAME :
  ------------------------------------------------------------
  :    63   : SHEUPDATA     :   142664   : SHAPEUPDB01   :
  ------------------------------------------------------------

  ------------------------------------------
  : INQSECTION NAME : TYPE : INQ FILE NO  :
  ------------------------------------------
  :   QSHEUPDATA    : 1    : 63          :
  ------------------------------------------

    INQ DATA BASE RETRIEVE START

ｺﾏﾝﾄﾞ A? RETR VAL EQ 'DICT    01A*'
  8269      RECORDS FOUND

ｺﾏﾝﾄﾞ A? DISP P-NAME VAL/L300

                .
                .
                .

-- 0076 ------------------------------------------------------------------------------------------------------------------------
 P-NAME  DICT
  VAL    DICT    01ABANDON
  VAL    DICT    02VT
  VAL    DICT    03放棄する

-- 0077 ------------------------------------------------------------------------------------------------------------------------
 P-NAME  DICT
  VAL    DICT    01ABANDON
  VAL    DICT    02VT
  VAL    DICT    03まかす

-- 0078 ------------------------------------------------------------------------------------------------------------------------
 P-NAME  DICT
  VAL    DICT    01ABANDON
  VAL    DICT    02VT
  VAL    DICT    03ひだねる

-- 0079 ------------------------------------------------------------------------------------------------------------------------
 P-NAME  DICT
  VAL    DICT    01ABANDON
  VAL    DICT    02VT
  VAL    DICT    03見捨てる

-- 0080 ------------------------------------------------------------------------------------------------------------------------
 P-NAME  DICT
  VAL    DICT    01ABANDON
  VAL    DICT    02VT
  VAL    DICT    03放棄する
```

```
  VAL     DICT     04棄，子などを
  VAL     DICT     05述
-- 0081 ----------------------------------------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT     01ABANDON
  VAL     DICT     02N
  VAL     DICT     03解放
-- 0082 ----------------------------------------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT     01ABANDON
  VAL     DICT     02N
  VAL     DICT     03気まま
-- 0083 ----------------------------------------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT     01ABANDONED
  VAL     DICT     02A
  VAL     DICT     03捨てられた
-- 0084 ----------------------------------------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT     01ABANDONED
  VAL     DICT     02A
  VAL     DICT     03捨てばちになった
-- 0085 ----------------------------------------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT     01ABANDONEE
  VAL     DICT     02N
  VAL     DICT     03被遺棄者
  VAL     DICT     05述
-- 0086 ----------------------------------------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT     01ABANDONEE
  VAL     DICT     02N
  VAL     DICT     03被委付者
-- 0087 ----------------------------------------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT     01ABANDONER
  VAL     DICT     02N
  VAL     DICT     03遺棄者
  VAL     DICT     05述
-- 0088 ----------------------------------------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT     01ABANDONER
  VAL     DICT     02N
  VAL     DICT     03委付者
-- 0089 ----------------------------------------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT     01ABANDONMENT
  VAL     DICT     02N
  VAL     DICT     03見捨てること
```

```
-- 0224 ----------------------------------------------------------------------------------------------------
P-NAME  DICT
 VAL    DICT    01HAIRY
 VAL    DICT    02A
 VAL    DICT    03灰色い

-- 0225 ----------------------------------------------------------------------------------------------------
P-NAME  DICT
 VAL    DICT    01HAITI
 VAL    DICT    02N
 VAL    DICT    03ハイチ

-- 0226 ----------------------------------------------------------------------------------------------------
P-NAME  DICT
 VAL    DICT    01HAITIAN
 VAL    DICT    02A
 VAL    DICT    03Ｈａｉｔｉの

-- 0227 ----------------------------------------------------------------------------------------------------
P-NAME  DICT
 VAL    DICT    01HAITIAN
 VAL    DICT    02N
 VAL    DICT    03Ｈａｉｔｉ人

-- 0228 ----------------------------------------------------------------------------------------------------
P-NAME  DICT
 VAL    DICT    01HAITIAN
 VAL    DICT    02N
 VAL    DICT    03Ｈａｉｔｉ■

-- 0229 ----------------------------------------------------------------------------------------------------
P-NAME  DICT
 VAL    DICT    01HAJJI
 VAL    DICT    02N
 VAL    DICT    03ｈａｄｊｉ

-- 0230 ----------------------------------------------------------------------------------------------------
P-NAME  DICT
 VAL    DICT    01HAKA
 VAL    DICT    02N
 VAL    DICT    03ハーカー

-- 0231 ----------------------------------------------------------------------------------------------------
P-NAME  DICT
 VAL    DICT    01HAKE
 VAL    DICT    02N
 VAL    DICT    03ナラの■
 VAL    DICT    05■

-- 0232 ----------------------------------------------------------------------------------------------------
P-NAME  DICT
 VAL    DICT    01HAKEEM,HAKIM
 VAL    DICT    02N
 VAL    DICT    03■人
 VAL    DICT    04■■■■の

-- 0233 ----------------------------------------------------------------------------------------------------
P-NAME  DICT
 VAL    DICT    01HAKEEM,HAKIM
```

```
  VAL     DICT    03xebec

-- 0039 ----------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT    01ZEBRA
  VAL     DICT    02N
  VAL     DICT    03シマウマ
  VAL     DICT    05動

-- 0040 ----------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT    01ZEBRAWOOD
  VAL     DICT    02N
  VAL     DICT    03縞のある低材を採る高木
  VAL     DICT    05植

-- 0041 ----------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT    01ZEBRAWOOD
  VAL     DICT    02N
  VAL     DICT    03それに似た各種の樹木

-- 0042 ----------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT    01ZEBU
  VAL     DICT    02N
  VAL     DICT    03ホウギュウ
  VAL     DICT    05動

-- 0043 ----------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT    01ZEBU
  VAL     DICT    02N
  VAL     DICT    03コブウシ

-- 0044 ----------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT    01ZECH.
  VAL     DICT    02N
  VAL     DICT    03Zechariah2

-- 0045 ----------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT    01ZECHARIAH
  VAL     DICT    02N
  VAL     DICT    03ゼカリヤ

-- 0046 ----------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT    01ZECHARIAH
  VAL     DICT    02N
  VAL     DICT    03「ゼカリヤ書」

-- 0047 ----------------------------------------------------------------------
 P-NAME   DICT
  VAL     DICT    01ZECHIN
  VAL     DICT    02N
  VAL     DICT    03sequin1
```

VAL	DICT	03	日本語訳
VAL	DICT	04	補足説明
VAL	DICT	05	分　　野

(2) 辞書データベース更新システムの概要

本実験で使用したＰＲＯＬＯＧ言語であるShape Up には，画面モードによるデータベース更新プログラムが用意されている。本実験では，辞書の更新をこのツールを利用して行った。次にその概要を示す。

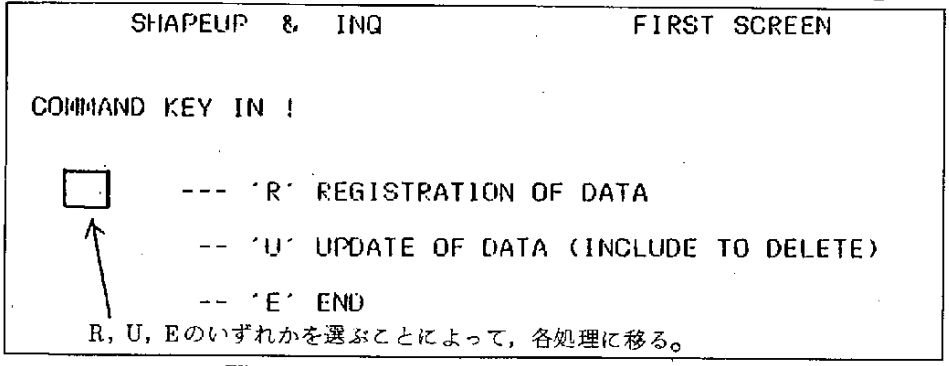


図 3.5－23　初 期 画 面

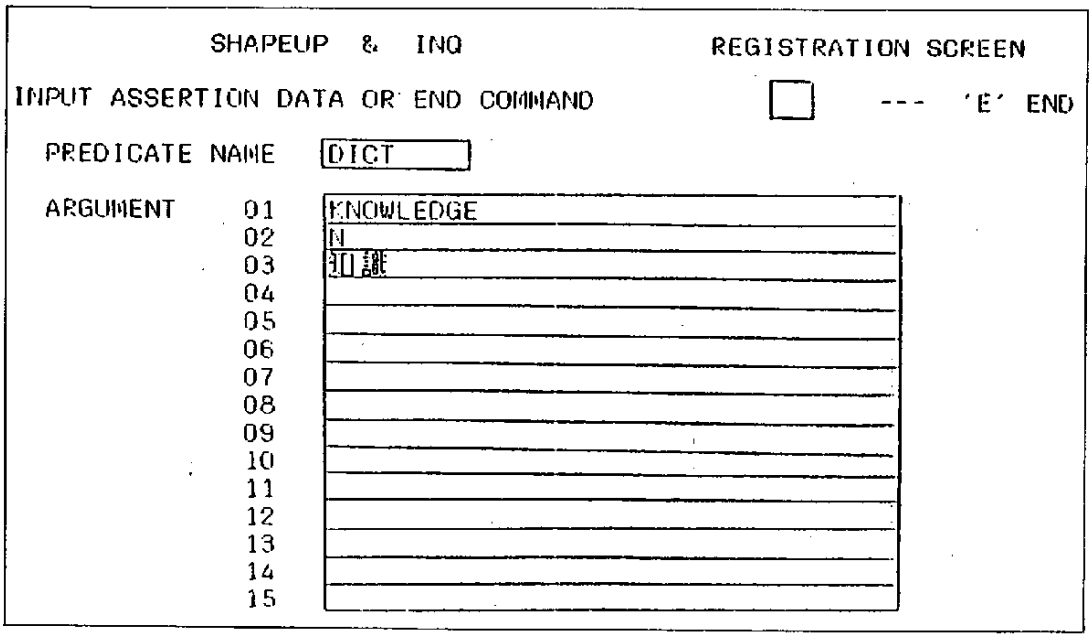


図 3.5－24　辞書データの登録

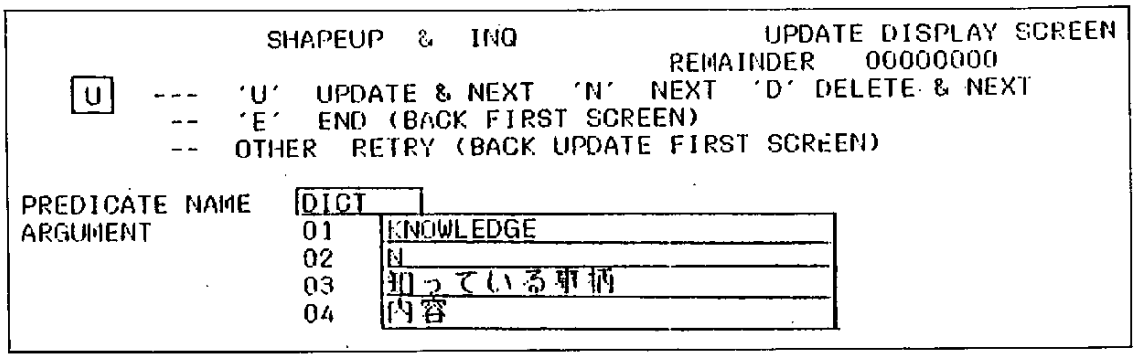


図3.5－25　辞書データの更新／削除

3.6　エネルギー分野における用語収集実験

(1)　実験の目的

今回の実験は自然言語処理に必要な辞書として，英日機械翻訳用の専門辞書及び検索効率を高めるためのシソーラス・ファイルを可能な限り機械的（人間による手作業を最小限に抑える方策）に作成することを目的に行った。

(2)　実験の概要

自然言語処理に必要な辞書類は，現在の技術水準からみて，対象分野を極力限定することにより実現することが最善の方策であるとの認識から，次のような分析フレームを設定した。

①　実験対象データ

米国ＥＩＣ社作製のエネルギー分野におけるアブストラクトデータで，１９８１年１月から１２月までの約5,０００記事の本文データを対象とする。

本データは，エネルギーの政策，資源，転換及び消費に関する科学的，工学的，政治的及び社会経済的観点からみた情報で，対象とする主題の範囲は，米国の経済・政策・計画等エネルギー一般の話題を始め，石油，

天然ガス，石炭，太陽エネルギー，原子力，核融合等のエネルギー種源別の情報やエネルギー生産，消費等の幅広い情報が各種雑誌，報告書，会議議事録等から作成されている。（表3.6－1参照）

② 処理概要

この実験データからアブストラクト部のみを抽出し，単語の切り出しを行った。

イ 記事数	4,999
ロ 切り出し単語数	422,515
（異なり語数	17,604）
ハ 一記事中の最大単語数	392
ニ 一単語の最大文字数	43

単語の切り出しを実際に行ってみて，いくつかの興味深い点が判明した。

第1は，エネルギー分野における専門用語の抽出を目的とするには，あまりにも単語レベルでは不充分なこと。

第2は，今回の実験で使用したデータでは，記事数が増えることに従い，出現する異なり語数がさらに増加傾向を示していることから，データ量不足であること。

第3は，単語切り出しのデリミタとしてスペースを採用した訳であるが，文末の単語についても正確に抽出しようとしたため単語文字列に連続したピリオド・スペースは分離することにした。この処理による弊害は次のような形で現れた。

U．S．△ → U．S△．△

これは，一つの記事データの中に複数の文データが入っている場合の文抽出方法の課題を示唆している。

第2，第3の問題は別途解決することして，第1の問題である単語レベルでは専門用語として不十分であるとの認識から再び用語抽出の課題

表3.6－1　エネルギーデータ

(1) AN: 01-81-20001
TI: (60) FINANCING THE FUTURE GROWTH OF THE ELECTRIC POWER INDUSTRY.
AU: (17) BAUGHMAN, M. L.
(27) KAHAL D. P. UNIV OF TEXAS,
SO: (38) NTIS REPORT UT/CES-PS-4, SEP 78 1901
ID: (24) 0ENERGY DEVEL FINANCING
(27) 0PUBLIC ELECTRIC UTILITIES
(15) 1CAPITAL COSTS
(19) 1ELECTRICITY RATES
(10) 2REVENUES
(24) 2INVESTMENT TAX CREDITS
AB: (499) SPECIAL REPORT
CAPITAL REQUIREMENTS AND FINANCIAL PROSPECTS OF THE U.S. ELECTRIC UTILITY INDUSTRY ARE ASSESSED WITHIN THE CONTEXT OF EXISTING REGULATORY INSTITUTIONS AND ELECTRICITY PRICES. MANY FINANCIAL DIFFICULTIES FOR THIS SECTOR MAY ARISE AS REGULATED PRICES ARE NOT SET HIGH ENOUGH TO ALLOW FOR SUFFICIENT CAPITAL TO BE GENERATED. ONE SOLUTION IS TO RAISE THE ALLOWED RATE OF RETURN TO A LEVEL GREATER THAN OR EQUAL TO THE COST OF CAPITAL. OTHER POLICIES FOR INCREASING REVENUES ARE SURVEYED.

(2) AN: 01-81-20002
TI: (49) RESEARCH INTO THE METHODOLOGY OF THE LEAP MODEL.
AU: (11) FALK, JAMES E. GEORGE WASHINGTON UNIV.
(18) MCCORMICK GARTH P.
(17) SOLAND RICHARD M.
SO: (40) DOE REPORT DOE/EIA-451001, DEC 79 (120)
ID: (25) 0ECONOMICS, ENERGY USAGE
(29) 0MATHEMATIC MODELS-ECONOMICS
(16) 0PRICES, ENERGY
(26) 1MATHEMATIC MODELS,ENERGY
AB: (719) SPECIAL REPORT
THE LEAP MODEL, WHICH DESCRIBES THE FLOW, WITH TIME, OF ENERGY RESOURCES AND THEIR PRICES OVER THE U.S. SUPPLY AND DEMAND NETWORK, IS ANALYZED. THE ECONOMIC MODEL IS MATHEMATICALLY A SIMULATION MODEL DESCRIBED BY SETS OF EQUATIONS GOVERNING RELATIONS BETWEEN QUANTITIES AND PRICES OF VARIOUS COMMODITIES AT VARIOUS PROCESSES OVER TIME. THE PRINCIPAL FEATURES OF THE LEAP MODEL ARE DESCRIBED. THE STRUCTURE OF LEAP'S EQUATIONS IS SUCH THAT THEIR SOLUTION (IF ONE EXISTS) CAN BE HANDLED WITH THE SUCCESSIVE OVER-RELAXATION METHOD. ASPECTS AND EXTENSIONS OF THIS METHOD, AND ALTERNATIVE METHODS COMMONLY USED TO SOLVE SIMULTANEOUS EQUATIONS ARE REVIEWED. (NUMEROUS DIAGRAMS, 13 REFERENCES, NUMEROUS TABLES)

(3) AN: 02-81-20003
TI: (48) ENERGY TAXATION: AN ANALYSIS OF SELECTED TAXES.
SO: (46) DOE REPORT DOE/EIA-0201/14, SEP 80, VIA (1069)
ID: (15) 0TAXATION, FED
(26) 0ENERGY USAGE, INDUSTRIAL
(13) 1PRICES, OIL
(23) 1TAXATION, STATE LOCAL
(17) 1OIL EXPLORATION
(21) 1OFFSHORE EXPLORATION
(21) 1ECONOMIC INCENTIVES
(17) 1COAL PRODUCTION
AB: (823) SPECIAL REPORT
THE EFFECTS OF VARIOUS TYPES OF ENERGY TAXES ON THE U.S. ECONOMY, ON ENERGY PRODUCTION AND UTILIZATION, AND ON SOCIAL INSTITUTIONS ARE EXAMINED. TAX TYPES CONSIDERED ARE ENERGY RESOURCE PRODUCTION AND PROPERTY TAXES, ENERGY USE AND CONSERVATION TAXES, TAX PREFERENCES, AND TAXES ON INCOME FROM ENERGY PRODUCTION. THE IMPLEMENTATION OF SUCH TAXES WOULD RAPIDLY INCREASE ENERGY PRICES, FACILITATE THE INTERACTION BETWEEN DIFFERENT ENERGY MARKETS, ENABLE A SMOOTH LONG-TERM TRANSITION TO ALTERNATIVE ENERGY SOURCES, AND ULTIMATELY PROVIDE FOR U.S. ENERGY SELF-SUFFICIENCY. IN ADDITION, ENERGY TAXES WILL INFLUENCE ENERGY DEVELOPMENT INVESTMENT DECISIONS THAT AFFECT THE ECONOMY'S PRODUCTIVITY AND RATE OF GROWTH. THE ENERGY TAX SYSTEM AS IT NOW EXISTS MUST BE REVISED. (1 GRAPH, 61 REFERENCES, NUMEROUS TABLES)

に取り組んだ。ここで一つの割り切りとして，連続した単語列が他の記事にも複数回出現すれば，それは用語として何らかの有意性があるとの判断から，最高１０単語構成から順次単語数を減らして出現状況を調べてみると表3.6－2のようになる。

但し，本処理の効率化を図るため，熟語・複合語等の先頭，中間，末尾等にあり得ない単語のリストアップを行い，これらをストップワードとして採用した。

なお，ストップワードの選定は以下のものを対象に選定した。

＜ストップワードの分類＞

- 冠　詞
- セパレータ
- 接続詞
- 前置詞
- 助動詞
- 代名詞
- 副　詞
- 動　詞（名詞，形容詞とならないもの

} 頻度１５以上から選定

- 動　詞（過去分詞）
- 形容詞

} 頻度９９以上から選定

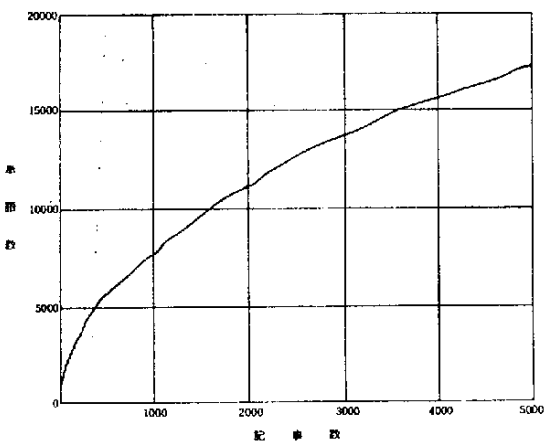


図3.6－1　記事数と単語数の関係

表3.6－2　熟語出現状況

熟語構成単語数	単語出現数	切出し熟語数	IDを付与した熟語数	頻度別熟語数 1	2	3	4	5	6	7以上	熟語選択(頻度)・(熟語数)	置換した熟語の総件数	熟語置換後の単・熟語出現数
10	422,515	17,595	17,498	17,425	61	8	0	1	2	1	2以上 73	123	421,408
9	421,408	22,271	22,220	22,201	12	2	3	1	0	1	2 〃 19	70	420,848
8	420,848	28,510	28,451	28,413	27	4	4	3	0	0	2 〃 38	97	420,169
7	420,169	36,056	35,937	35,857	60	12	4	2	1	1	2 〃 80	199	418,975
6	418,975	45,195	44,970	44,801	136	20	7	4	0	2	2 〃 169	394	417,005
5	417,005	57,096	56,489	56,055	338	59	20	7	3	7	2 〃 434	1,037	412,857
4	412,857	69,110	66,852	65,437	1,069	180	63	41	18	44	3 〃 346	1,534	408,255
3	408,255	76,894	67,687	62,924	3,228	767	296	149	77	246	4 〃 768	5,202	397,851
2	397,851	91,777	57,774	46,783	5,925	2,070	952	541	307	1,196	5 〃 2,044	22,539	375,312

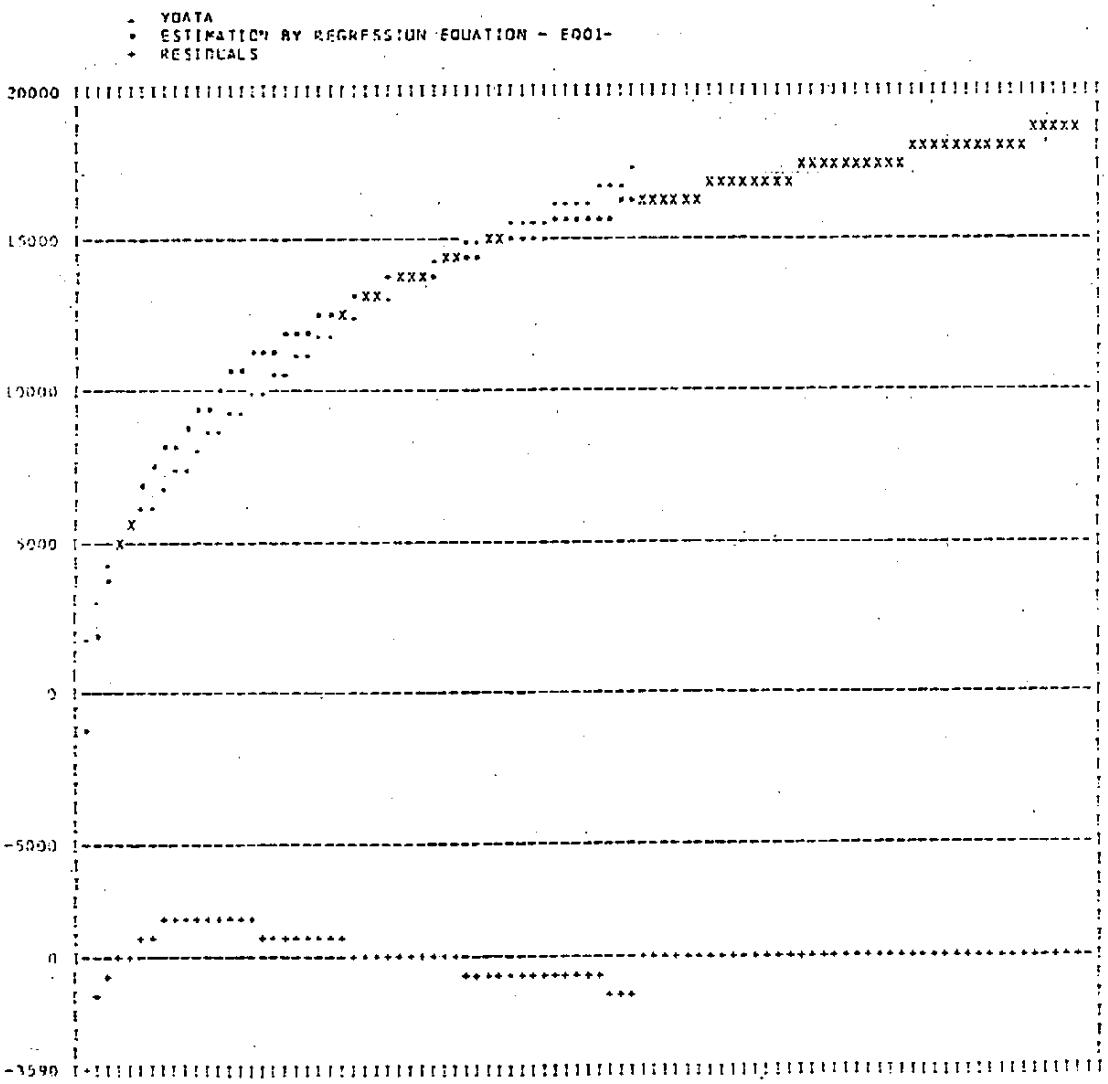
熟語選択範囲　（計3,971）

熟語処理後

切出し単・熟語数	375,312 語
IDを付与した単語数	21,376 語
頻度表（頻度6以上）	6,370 語

（頻度）

1	2	3	4	5	6	7以上
7,986語	2,919語,	1,522語,	1,243語,	1,335語,	854語,	5,516語



$Y = a + b \log (\times)$

単語数　　記事数

回帰結果

$Y = -21974.9 + 10280.9 \log (\times)$

決定係数　0.94

重相関係数　0.97

図 3.6－2　記事数と新単語出現との関係

熟語・複合語の候補抽出結果を表3.6－3に示す。この結果から，10単語のように長い単語から構成される用語は，時代の重要なトピックが現れているのがわかる。また辞書に収録すべき用語は2～4単語構成を中心に考えるのが効果的であることもわかる。

次に熟語・複合語を含めた語の出現頻度状況を調べると，表3.6－4のようになる。高頻度リストに登場する語は，いわゆるコモンワードであり，本実験の目的とするエネルギー分野の専門用語は中頻度位置に比較的多く出現している。

なお，低頻度語は全体の半分を占め，これらはむしろ特殊語として位置付けることができる。

従って，以上のことを踏まえ，用語の頻度分析を行うと，概念的には図3.6－3のように考えることができる。

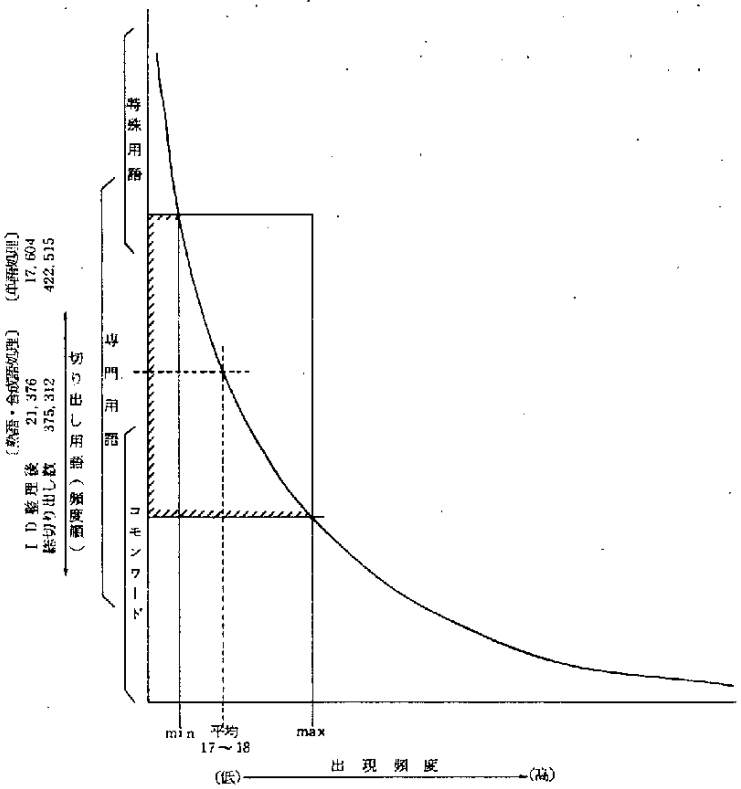


図3.6－3　　頻度分析

表3.6－3　熟語・複合語抽出リスト
①頻度数　10単語構成

1	951110	61	AVERAGE PRICE OF DOMESTIC CRUDE OIL PURCHASED AT THE WELLHEA U	7	7	962	973	1499	1509	2512	3579	4559	
2	951115	54	AVERAGE RETAIL PRICE FOR ALL GRADES AND TYPES OF MOTOR	6	6	962	973	1499	1509	2512	3579		
3	963035	55	RETAIL PRICE FOR ALL GRADES AND TYPES OF MOTOR GASOLINE	6	6	962	973	1499	1509	2512	3579		
4	950078	58	ACCIDENT AT THE THREE MILE ISLAND NUCLEAR POWER PLANT NEAR	5	5	322	348	835	874	882			
5	950950	68	ASSISTANCE TO STATE AND LOCAL GOVERNMENTS TO PROMOTE NATIONA L ENERGY	3	3	2039	2995	2996					
6	963836	68	SALES OF ELECTRICITY TO ULTIMATE CONSUMERS BY ALL ELECTRIC U TILITIES	3	3	387	919	2015					
7	967360	49	WIND TUNNEL STUDY OF A SERIES OF MODEL FLAT PLATE	3	3	4887	4888	4889					
8	965129	53	STUDY OF A SERIES OF MODEL FLAT PLATE SOLAR COLLECTOR	3	3	4887	4888	4889					
9	959757	66	MODEL FLAT PLATE SOLAR COLLECTOR INSTALLATIONS IMMERSED IN A THICK	3	3	4887	4888	4889					
10	964498	71	SOLAR COLLECTOR INSTALLATIONS IMMERSED IN A THICK TURBULENT SHEAR LAYER	3	3	4887	4888	4889					
11	956425	70	FLAT PLATE SOLAR COLLECTOR INSTALLATIONS IMMERSED IN A THICK TURBULENT	3	3	4887	4888	4889					
12	961516	71	PLATE SOLAR COLLECTOR INSTALLATIONS IMMERSED IN A THICK TURB ULENT SHEAR	3	3	4887	4888	4845					
13	955230	79	ENERGY CONVERSION SYSTEMS AND ASSOCIATED AUXILIARIES OF GEOT HERMAL POWER PLANTS	2	2	2375	2394						
14	955270	71	ENERGY EFFICIENT PASSIVE SOLAR HOUSING FOR THE ELDERLY IN MA SSACHUSETTS	2	2	1383	3972						
15	956167	55	FEA AS MANDATED BY THE ENERGY POLICY & CONSERVATION ACT	2	2	924	925						
16	958175	72	FEASIBILITY OF ADVANCED CONCEPTS FOR GENERATING PEAK LOAD EL ECTRIC POWER	2	2	2375	2394						
17	956316	67	FINANCE INCREMENTAL COSTS OF PASSIVE SOLAR DESIGNS IN LOCAL HOUSING	2	2	1383	3972						
18	956325	80	FINANCIAL ASSISTANCE FOR THE DEVELOPMENT OF COMMERCIAL SYNTH ETIC FUEL PRODUCTION	2	2	2641	2642						
19	952408	83	COMPREHENSIVE CONSERVATION PROGRAM IMPLEMENTED BY ANY STATE OR MUNICIPAL GOVERNMENT	2	2	408	950						
20	952421	82	COMPRESSED AIR ENERGY STORAGE POWER PLANT INCORPORATING A CO AL-FIRED FLUIDIZED-BED	2	2	2375	2394						
21	952491	66	CONCEPTS FOR GENERATING PEAK LOAD ELECTRIC POWER FROM A COMP RESSED	2	2	2375	2394						
22	956460	64	GENERATING PEAK LOAD ELECTRIC POWER FROM A COMPRESSED AIR EN ERGY	2	2	2375	2394						
23	957099	59	GOVERNMENTS CAN TAKE TO ALLEVIATE THE BURDEN OF HIGH ENERGY	2	2	2102	2103						
24	957347	72	HEATING AND COOLING RETROFIT OPTIONS CAN SIGNIFICANTLY REDUC E THE ENERGY	2	2	3814	4980						
25	957366	66	HEATING OIL SALES BY REFINERS TO RESELLERS AND RETAILERS ING REASED	2	2	916	927						
26	957410	54	HELD TO CONSIDER THE IMPACT OF ENERGY COSTS ON ELDERLY	2	2	2102	2103						
27	957411	62	HELD TO CONSIDER THE PASSAGE OF THE PROPOSED ENERGY MANAGEME NT	2	2	2955	2956						
28	957413	56	HELD TO DISCUSS THE CRUDE OIL WINDFALL PROFITS TAX ACT	2	2	1060	1062						
29	958855	64	IMPROVED BY THE APPLICATION OF A TRANSPARENT HEAT MIRROR COA TING	2	2	1327	1315						
30	958607	55	INTERFACE BETWEEN A SOLAR PHOTOVOLTAIC ARRAY AND AN AC LOAD	2	2	126	127						
31	956550	81	IRRADIANCE AND TEMPERATURE MEASUREMENTS IN CONJUNCTION WITH SOLAR SPECTRUM MODELS	2	2	1763	1820						
32	958558	62	LOAD ELECTRIC POWER FROM A COMPRESSED AIR ENERGY STORAGE PLA NT	2	2	2375	2394						
33	959297	72	MARINE BUNKERS AND DELIVERIES TO INTERNATIONAL CIVIL AVIATIO N BY PRODUCT	2	2	3494	3495						

② 頻度数　4 単語構成

1/7

1	456641	16	STATE OF THE ART	25	20	113	362	470	677	626	1133	1192	1687
2	440067	19	OIL AND NATURAL GAS	21	20	13	416	426	431	711	913	1434	1511
3	435903	31	MEMORANDA AND RELATED DOCUMENTS	21	21	260	1060	1061	1109	1427	1476	1960	2102
4	400083	27	CAPITAL AND OPERATING COSTS	16	15	343	434	440	495	627	652	711	3279
5	460304	30	WIND ENERGY CONVERSION SYSTEMS	16	15	1650	1651	1666	1667	2757	2763	2766	3777
6	420105	12	ENERGY R & D	14	13	11	526	1578	2147	2675	3001	3202	3603
7	440047	22	OIL AND GAS PRODUCTION	13	13	46	1089	2612	2705	3163	3190	3703	3711
8	448356	14	R & D PROGRAMS	11	11	282	673	694	705	882	951	1843	2143
9	455403	25	SOLAR HEATING AND COOLING	11	11	705	1751	2111	2249	2305	3314	3793	3669
10	459758	34	TECHNICAL AND ECONOMIC FEASIBILITY	11	11	420	728	1051	1630	2391	2774	2959	2962
11	417025	29	DOMESTIC CRUDE OIL PRODUCTION	10	10	958	969	1495	1505	2500	2991	3221	3575
12	455666	27	SOLAR WATER HEATING SYSTEMS	10	9	144	1405	1420	2273	2282	2283	2848	3912
13	414391	25	DEPENDENCE ON FOREIGN OIL	9	9	6	277	275	1428	2022	2125	2141	3690
14	423887	27	FLAT PLATE SOLAR COLLECTORS	9	9	1345	1469	1771	1872	1895	2799	3825	4909
15	438111	22	NATURAL GAS POLICY ACT	9	9	435	902	1562	1615	2100	2649	3160	4720
16	453180	28	SECOND LAW OF THERMODYNAMICS	9	9	56	191	362	369	409	412	444	447
17	448317	16	R & D ACTIVITIES	9	9	622	633	2295	2693	2556	3367	4095	4222
18	448328	13	R & D EFFORTS	9	9	69	106	148	881	996	1640	3273	3500
19	440052	21	OIL AND GAS RESOURCES	9	9	573	2192	2688	2654	2698	3194	3221	4221
20	449949	17	REGION OF THE U.S	9	9	558	559	561	562	668	782	1643	3216
21	450007	18	REGIONS OF THE U.S	9	9	1263	1353	1521	1730	2151	2856	2989	3056
22	440029	23	OIL AND GAS EXPLORATION	8	7	4	25	542	565	2154	2710	3713	
23	440035	20	OIL AND GAS INDUSTRY	8	7	25	530	1025	2150	2701	3702	4265	
24	448347	14	R & D PROJECTS	8	8	622	660	700	1525	2045	4368	4372	5073
25	456597	27	STATE AND LOCAL GOVERNMENTS	8	7	2020	2116	2130	2405	2419	2876	5201	
26	460099	26	THREE MILE ISLAND ACCIDENT	8	6	276	306	320	875	2471	5111		
27	423922	27	FLORIDA SOLAR ENERGY CENTER	8	8	1795	2259	2269	3873	3911	3964	4028	4892
28	409674	29	COMPRESSED AIR ENERGY STORAGE	8	8	2366	2371	2374	2393	2395	2396	2397	2398
29	427679	25	HEATING AND COOLING LOADS	8	7	315	3893	3920	3917	4015	4332	4987	
30	414397	26	DEPENDENCE ON IMPORTED OIL	7	7	423	1450	2968	3660	4232	4451	5209	
31	423886	26	FLAT PLATE SOLAR COLLECTOR	7	7	1420	1900	1901	1905	1506	2239	4980	
32	423769	23	FIRST YEAR OF OPERATION	7	6	58	2860	3305	3995	4134	4336		
33	452170	24	SALT GRADIENT SOLAR POND	7	7	118	1886	2922	3923	3929	4385	4386	
34	463114	20	USE OF ENERGY MODELS	7	6	537	539	540	542	818	934		
35	463192	20	USE OF NUCLEAR POWER	7	7	498	536	808	2608	2669	2982	1065	
36	455569	27	SOLAR SPACE HEATING SYSTEMS	7	7	1786	1867	1871	2880	2881	3897	4344	
37	405443	18	BILLION BBL OF OIL	7	7	47	85	579	1626	2159	4266	4267	
38	448345	13	R & D PROGRAM	7	7	1136	1411	1739	2296	2474	3446	3476	
39	448824	15	RD & D PROJECTS	7	6	147	1547	1593	1829	2188	2676		
40	456113	23	SPACE AND WATER HEATING	7	7	139	1543	1547	2386	3929	3967	4368	
41	456124	25	SPACE HEATING AND COOLING	7	7	1128	1347	1365	3258	3336	4019	4597	
42	461293	32	TOTAL PETROLEUM IMPORTS AVERAGED	7	7	958	969	1495	1505	2500	3575	4555	
43	442362	29	PASSIVE SOLAR HEATING SYSTEMS	7	7	1214	1329	1720	2859	3622	3938	4957	
44	441022	23	OPERATION OF THE SYSTEM	7	7	219	1217	1350	2317	3510	3860	4505	
45	442305	27	PASSIVE HEATING AND COOLING	6	6	1701	2594	4357	4966	4568	4981		
46	439386	26	NUCLEAR POWER PLANT SAFETY	6	6	318	327	3431	3438	3446	3481		
47	445671	24	PRICES OF MOTOR GASOLINE	6	6	962	973	1509	1512	3579	4559		
48	446665	27	PRODUCTION OF ALCOHOL FUELS	6	6	691	2109	2632	2633	3750	4374		
49	455397	23	SOLAR HEATING / COOLING	6	6	134	138	145	147	3181	3900		
50	457125	23	STORAGE OF SOLAR ENERGY	6	6	121	2291	3826	4334	4435	4912		
51	459820	37	TECHNICALLY AND ECONOMICALLY FEASIBLE	6	6	1453	1966	2755	2959	2961	4313		
52	455650	26	SOLAR TOTAL ENERGY SYSTEMS	6	5	727	728	729	1916	4367			
53	408041	18	COAL AND OIL SHALE	6	5	25	2329	2550	4405	5021			
54	452744	20	SALES OF NATURAL GAS	6	6	959	1496	1506	2505	2989	3576		
55	415023	28	DEVELOPMENT OF NUCLEAR POWER	6	5	160	343	3529	3632	3681			
56	417063	26	DOMESTIC HOT WATER HEATING	[illegible]	[illegible]	[illegible]							

③ 頻度数　2単語構成

1	232447	11	NATURAL GAS	230	187	14	16	33	38	39	40	65	156
2	247505	12	SOLAR ENERGY	219	183	19	27	31	107	111	114	118	132
3	215221	19	ENERGY CONSERVATION	181	157	27	366	366	381	382	392	406	417
4	215226	18	ENERGY CONSUMPTION	132	111	202	375	402	403	444	448	453	472
5	233873	13	NUCLEAR POWER	122	98	253	257	258	260	262	272	269	297
6	210415	9	CRUDE OIL	121	95	8	17	26	37	38	40	83	166
7	215468	13	ENERGY POLICY	89	70	13	17	42	275	277	366	498	682
8	230874	11	MILLION BPD	89	44	17	52	204	390	573	921	953	958
9	230581	11	POWER PLANT	86	81	147	233	235	276	282	335	407	456
10	257142	11	WIND ENERGY	86	67	19	27	28	80	90	134	137	147
11	215628	10	ENERGY USE	84	70	3	9	376	381	479	483	915	1014
12	202787	15	BEING DEVELOPED	80	79	94	134	143	144	147	150	249	268
13	215521	19	ENERGY REQUIREMENTS	78	73	19	28	150	153	439	887	1077	1302
14	244068	16	RESULTS INDICATE	77	77	134	467	713	730	754	824	863	1082
15	234496	9	OIL SHALE	75	58	72	85	86	89	173	179	180	182
16	233020	11	NEW ZEALAND	74	63	101	1097	1547	1613	2178	2271	2783	2787
17	215526	16	ENERGY RESOURCES	73	65	2	36	85	89	374	395	412	444
18	222315	9	HEAT PUMP	73	50	105	114	458	478	491	733	736	1271
19	252085	19	THERMAL PERFORMANCE	72	65	441	491	724	730	1173	1195	1201	1237
20	223200	9	HOT WATER	68	58	66	129	377	478	488	493	602	644
21	230858	12	MILLION TONS	66	32	58	176	177	386	391	763	952	960
22	233830	14	NUCLEAR ENERGY	66	54	10	17	20	275	277	293	317	359
23	215539	14	ENERGY SAVINGS	66	56	431	436	457	479	481	487	759	926
24	214590	14	ELECTRIC POWER	66	59	27	257	298	403	479	684	688	723
25	215553	13	ENERGY SUPPLY	64	60	84	189	317	422	527	593	666	682
26	248171	13	SPACE HEATING	64	58	73	101	115	377	458	478	482	488
27	215265	13	ENERGY DEMAND	61	47	368	457	483	494	548	550	683	687
28	230612	11	MILLION BBL	60	38	230	386	389	390	391	564	920	971
29	247643	15	SOLAR RADIATION	57	52	122	151	358	1011	1186	1188	1238	1302
30	212532	16	DISTRICT HEATING	58	35	33	378	566	567	644	732	2219	2279
31	217562	18	FEDERAL GOVERNMENT	57	54	11	28	45	89	209	353	404	416
32	215293	17	ENERGY EFFICIENCY	56	47	415	433	448	485	490	549	1256	1544
33	234457	14	OIL PRODUCTION	55	50	44	52	53	140	183	220	941	969
34	230583	12	POWER PLANTS	53	52	63	137	207	235	242	263	283	300
35	250787	15	SYNTHETIC FUELS	53	48	16	17	172	186	204	528	537	543
36	214605	18	ELECTRIC UTILITIES	53	51	198	236	373	384	419	423	679	817
37	215275	18	ENERGY DEVELOPMENT	52	46	3	7	10	17	33	363	512	514
38	247697	13	SOLAR SYSTEMS	51	50	102	139	713	728	755	1191	1261	1297
39	223570	19	HYDROELECTRIC POWER	50	45	19	27	28	253	370	498	558	560
40	215137	12	FOSSIL FUELS	48	45	70	139	159	179	386	391	396	419
41	222339	12	HEAT STORAGE	48	36	108	454	486	1213	1214	1236	1245	1261
42	234349	13	OIL COMPANIES	48	41	6	8	16	1063	1088	1089	1095	1105
43	212612	11	DIRECT GAIN	48	47	1157	1161	1164	1165	1169	1196	1198	1202
44	215592	15	ENERGY SUPPLIES	47	45	18	42	543	657	660	666	740	1128
45	247553	13	SOLAR HEATING	47	39	1157	1166	1205	1216	1221	1310	1605	1710
46	222325	13	HEAT RECOVERY	47	39	67	125	434	703	1118	1943	2046	2149
47	222348	13	HEAT TRANSFER	46	35	408	413	720	734	738	753	1189	1239
48	250973	18	SYSTEM PERFORMANCE	46	45	723	1164	1175	1182	1217	1294	1334	1340
49	243065	17	RELATED DOCUMENTS	46	46	392	393	1057	1058	1059	1062	1063	1434
50	215435	12	ENERGY NEEDS	46	42	14	29	34	152	228	532	706	975
51	215600	14	ENERGY SYSTEMS	45	43	143	152	191	369	372	441	478	496
52	215571	13	ENERGY SOURCE	45	44	67	80	92	390	385	416	441	590
53	252119	15	THERMAL STORAGE	46	44	248	250	1165	1168	1207	1225	1226	1235
54	221201	20	GEOTHERMAL RESOURCES	45	38	96	557	558	710	1128	1608	1670	2172
55	247650	16	SOLAR COLLECTORS	44	43	165	836	851	858	721	1240	1293	1313
56	215572	14	ENERGY SOURCES	43	40	10	20	152	257	492	646	1752	[illegible]

表3.6－4　頻度リスト　①高頻度例

1	10	1	.	21083	4999	1	2	3	4	5	6	7	8	
2	117690	3	THE	21247	4035	1	2	3	4	5	6	7	8	
3	190	1	,	10269	4325	2	3	5	6	7	8	9	10	
4	119550	2	OF	14734	4654	1	2	3	4	5	6	7	8	
5	7700	3	AND	13678	4685	1	2	3	5	6	7	9	10	
6	100270	2	TO	7725	3711	1	2	3	4	5	6	7	10	
7	86900	2	IN	7382	3676	3	4	7	8	9	10	11	16	
8	9740	3	ARE	6818	3755	1	2	3	4	5	6	7	8	
9	670	1	A	6229	3198	1	2	3	6	8	10	11	12	
10	93520	2	IS	5075	3047	1	2	5	8	9	12	14	15	
11	49830	3	FOR	5021	3035	1	3	4	5	6	7	8	10	
12	14540	2	BE	2323	1701	1	2	3	4	6	7	8	10	
13	177660	4	THAT	2221	1745	2	3	4	8	11	13	20	21	
14	20780	2	BY	2114	1600	2	7	8	10	11	12	13	18	
15	196650	4	WITH	2036	1562	2	11	14	16	18	20	22	23	
16	121250	2	ON	1792	1375	3	5	6	8	11	12	18	20	
17	10360	2	AS	1634	1233	1	3	8	14	18	24	30	32	
18	71770	4	FROM	1571	1218	3	4	9	12	17	24	28	30	
19	100	1	)	1533	615	7	27	30	41	49	51	56	65	
20	7300	2	AN	1465	1202	8	12	14	21	22	32	34	35	
21	192670	3	WAS	1293	789	22	32	43	50	56	59	71	73	
22	194550	4	WERE	1288	871	7	25	37	42	43	56	73	75	
23	11300	2	AT	1077	806	2	14	24	25	27	31	38	39	
24	200	1	%	1010	658	4	7	19	20	25	28	35	38	
25	176610	5	THESE	1004	877	4	6	7	10	11	13	24	25	
26	40100	9	DISCUSSED	996	979	5	6	15	16	29	37	38	44	
27	44260	9	DESCRIBED	981	969	2	24	28	34	35	43	44	47	
28	178870	4	THIS	973	864	1	2	11	19	29	30	37	40	
29	105060	3	U.S	967	716	1	2	3	6	10	12	13	14	
30	79180	3	HAS	893	753	4	8	11	14	21	24	26	30	
31	79260	4	HAVE	881	730	8	10	17	18	21	41	42	46	
32	4330	4	ALSO	813	774	5	7	11	14	17	18	22	44	
33	174510	6	SYSTEM	807	584	3	23	74	82	93	94	99	105	
34	21310	3	CAN	757	650	2	9	10	18	41	52	56	71	
35	122690	2	OR	723	617	1	4	5	7	11	14	32	42	
36	56910	6	ENERGY	713	505	3	10	16	19	20	31	56	61	
37	52000	6	DURING	696	493	6	8	9	10	24	26	40	53	
38	204040	4	1980	683	324	4	31	46	50	56	176	177	178	
39	14040	4	BEEN	672	575	4	8	11	17	30	42	52	53	
40	195410	4	WILL	652	515	20	21	22	23	25	26	36	40	
41	174760	7	SYSTEMS	652	526	18	67	74	97	98	102	115	116	
42	100330	3	USE	647	518	9	34	36	37	46	57	67	77	
43	171440	4	SUCH	627	503	2	3	18	30	32	50	63	74	
44	162460	5	SOLAR	619	468	7	16	28	107	109	112	115	116	
45	177650	4	THAN	611	510	1	8	9	12	14	26	28	29	
46	180	1	/	552	400	5	71	76	84	110	125	130	163	
47	45590	11	DEVELOPMENT	552	482	12	13	17	21	30	33	44	45	
48	62030	8	EXAMINED	549	542	3	6	8	12	15	20	28	45	
49	27440	4	COAL	539	344	7	10	16	27	56	58	59	61	
50	130030	7	PROGRAM	525	395	17	34	44	85	103	102	100	100	
51	100430	4	USED	523	403	28	32	34	47	60	71	75	76	
52	44400	6	DESIGN	515	442	67	69	71	110	117	120	154	160	
53	33730	10	CONSIDERED	493	402	3	11	15	62	66	67	83	92	
54	111220	4	MORE	492	427	4	5	7	8	9	14	20	29	
55	137360	10	PRODUCTION	472	306	3	16	28	38	39	41	42	48	
56	81430	7	INCLUDE	469	450	13	17	27	31	47	59	67	90	

②中頻度例

2241	91330	5	INTENSITY	18	17	370	400	442	445	743	820	945	1150
2242	91690	9	INTERESTS	18	18	23	505	1250	1501	1619	1634	1787	2193
2243	110260	15	NUCLEAR REACTOR	18	17	269	837	840	844	86[illegible]	863	873	881
2244	97000	4	LASL	18	16	1170	1188	1197	1201	1220	1252	1383	1460
2245	97460	2	LD	18	14	259	367	679	766	1936	1963	1565	2379
2246	78020	5	GUIDE	18	17	240	314	837	1116	1514	2063	2357	2748
2247	76930	11	GOVERNMENTS	18	18	33	244	566	2136	2265	2271	2343	2643
2248	53370	9	ECONOMIES	18	16	395	406	435	800	993	1474	1524	2135
2249	54900	23	ELECTRICITY CONSUMPTION	18	16	410	455	494	1339	2037	3071	3341	3351
2250	54910	18	ELECTRICITY DEMAND	18	12	830	2021	2410	2487	2478	2589	3052	3509
2251	73850	3	GAO	18	16	541	1428	1989	2123	2410	2600	3106	3163
2252	74800	6	GEARED	18	18	592	633	673	904	904	1254	1345	1752
2253	55950	8	EMERGING	18	18	1461	1600	1605	1764	1771	2009	2671	2887
2254	58160	9	EMPIRICAL	18	17	368	490	526	1451	1779	1992	2535	2589
2255	57910	15	ENERGY PLANNING	18	15	1504	1592	2110	2451	2988	3005	3100	4474
2256	72070	10	FUEL CYCLE	18	16	257	349	352	362	869	872	1449	2411
2257	51630	7	DRIVING	18	15	462	473	1111	1632	2053	2539	2547	2565
2258	81310	15	HIGH EFFICIENCY	18	18	736	822	988	1144	1213	1271	1755	2197
2259	65470	7	FARMING	18	13	670	689	694	1522	1628	3280	3342	3754
2260	50450	22	ENERGY STORAGE SYSTEMS	18	17	248	470	478	1536	2368	2369	2370	2378
2261	58570	17	ENERGY TECHNOLOGY	18	18	660	1301	1636	2146	2524	2902	3003	3554
2262	59700	8	ENVELOPE	18	11	753	754	1150	1242	1246	1330	3040	3966
2263	60670	5	ERROR	18	13	311	326	328	850	860	1129	1276	1731
2264	62760	5	EXIST	18	17	4	40	54	207	259	580	2422	3661
2265	125160	4	PAGE	18	18	49	50	64	553	1025	1608	1612	1661
2266	136330	14	PRIMARY ENERGY	18	17	20	450	682	803	885	1480	1543	2227
2267	145840	7	REFLECT	18	18	581	758	947	1484	2018	2037	2111	2249
2268	129200	18	PETROLEUM INDUSTRY	18	13	1621	2033	2104	2161	3169	3186	3226	3534
2269	130460	11	PILOT PLANT	18	16	162	173	531	630	768	1407	1925	1926
2270	130180	19	PHOTOVOLTAIC SYSTEM	18	15	728	1749	2123	3312	3316	3320	3695	3982
2271	131360	6	PLASMA	18	12	902	904	1144	1956	2009	2295	3395	5131
2272	148130	6	REPAIR	18	11	309	719	836	2800	3092	3390	3393	3473
2273	148980	17	RESEARCH PROGRAMS	18	18	250	310	1746	1891	2248	2568	2848	3195
2274	149450	8	RESIDUAL	18	17	180	198	780	1021	1457	1514	1947	1968
2275	126610	20	PASSIVE SOLAR DESIGN	18	15	1108	1244	1251	1257	1371	1714	1736	2289
2276	155120	9	SCHEDULED	18	18	44	123	204	1625	1671	1843	1976	2112
2277	155600	3	SEA	18	18	565	1407	1621	2071	2073	2155	2161	2721
2278	156040	7	SECTION	18	17	336	1330	1371	2063	2194	2536	2627	2773
2279	175520	9	TAR SANDS	18	15	948	1058	1474	1668	2036	2054	2321	2332
2280	174130	8	SYSTEMS	18	17	94	355	2036	2136	2795	2913	3029	3044
2281	163320	11	SOLAR HOUSE	18	16	1239	1246	1271	1416	1700	1701	1738	1859
2282	162260	7	SOCIETY	18	16	31	50	381	731	1274	1479	1585	1605
2283	197310	9	WORKSHOPS	18	17	990	1293	1296	1297	1297	1927	1715	2268
2284	178930	7	VIBRATION	18	9	259	261	349	352	557	872	1466	2414
2285	168300	15	STEAM GENERATOR	18	16	97	125	271	814	830	1942	3322	3409
2286	169170	16	STORAGE CAPACITY	18	17	406	707	811	1168	1236	1309	1484	1789
2287	187750	5	UPPER	18	16	407	1240	1313	1474	1599	2201	3278	3280
2288	186730	5	UNION	18	18	231	1058	1062	1427	1538	1942	2143	2159
2289	193240	12	WATER SUPPLY	18	16	907	1638	1641	1849	2075	2926	3112	3255
2290	197410	9	WORLD OIL	18	16	30	1073	1087	1088	2013	2137	2138	2517
2291	174590	13	SYSTEM DESIGN	18	18	410	955	1194	1760	1777	2030	2368	2757
2292	191390	4	VIEW	18	18	983	1174	1522	1776	2112	2119	2127	2476
2293	1020	9	ACCOUNTED	18	14	910	1073	1103	1184	2039	2154	3096	3571
2294	2780	9	ADDITIONS	9	15	1624	1767	2142	2488	2481	2925	2992	3685
2295	11200	6	ASSURE	3	18	55	262	814	1409	1577	1795	1988	2183
2296	11130	7	ASSURED	18	15	305	494	1089	1135	1226	1408	1721	1744

③低頻度例

・・・ヒットリスト・・・

	ＩＩＤＩ		ターゲット	ヒット								
6329	4450	23	AGRICULTURAL PRODUCTION	6	6	435	2072	2075	3591	3747	4606	
6330	4460	21	AGRICULTURAL RESIDUES	6	6	692	693	704	1015	2320	4245	
6331	12760	54	AVERAGE RETAIL PRICE FOR ALL GRADES AND TYPES OF MOTOR	6	6	962	973	1499	1509	2512	3519	
6332	134220	17	SYNFUELS INDUSTRY	6	6	545	1041	1432	2120	2128	4729	
6333	173130	13	SUPPLY SYSTEM	6	6	273	395	1454	2137	2294	3459	
6334	173240	10	SUPPORTIVE	6	6	1744	2820	3484	4024	4312	4511	
6335	172360	6	SUNCAT	6	6	1170	1171	1707	1784	3991	4947	
6336	172680	15	SUPERCONDUCTING	6	5	1444	2368	2399	4929	5140		
6337	192550	7	WARMING	6	6	394	499	1011	1128	1670	2301	
6338	191540	9	VOLUNTARY	6	6	317	2364	3099	3593	4030	5201	
6339	192840	10	WASTE FUEL	6	6	2006	3758	3759	4150	4291	4326	
6340	192920	23	WASTE-TO-ENERGY SYSTEMS	6	5	604	655	1117	1110	2190		
6341	193180	14	WATER REACTORS	6	3	1997	5112	5124				
6342	193200	12	WATER RIGHTS	6	4	2349	2952	3594	3649			
6343	190540	10	VEHICLE PROPULSION	6	6	465	670	1534	1535	2057	4654	
6344	190290	15	VARIOUS REGIONS	6	5	1714	2150	2485	4830	4091		
6345	190250	15	VARIOUS FEDERAL	6	6	1615	2121	3023	3035	3205	3347	
6346	190230	24	VARIOUS ENERGY RESOURCES	6	6	369	1075	2070	2498	3195	4219	
6347	190360	11	VARIOUS U.S	6	6	1475	2864	3169	4130	4233	4246	
6348	190310	15	VARIOUS SYSTEMS	6	6	251	299	370	515	1730	1803	
6349	191350	8	VIGOROUS	6	6	13	1073	1102	1474	1590	4726	
6350	191150	18	VIABLE ALTERNATIVE	6	6	1125	1775	1056	2911	3246	4203	
6351	191750	10	VOLATILITY	6	6	367	1668	2134	2580	3618	4649	
6352	181950	12	TRANSACTIONS	6	3	912	3478	5056				
6353	182220	15	TRANSIT SYSTEMS	6	5	1001	1532	1533	3099	3622		
6354	182330	19	TRANSMISSION SYSTEM	6	6	800	2367	3420	3421	3607	4896	
6355	194630	18	WESTERN HEMISPHERE	6	4	1120	1123	3691	4224			
6356	194640	14	WESTERN STATES	6	5	63	1023	1980	3174	4693		
6357	194150	16	WEIGHT REDUCTION	6	4	3053	3090	3091	3623			
6358	194790	5	WHEAT	6	6	2207	2322	2727	2759	3242	4810	
6359	194870	7	WHEREAS	6	6	45	156	332	1420	2134	2221	
6360	193230	14	WATER SUPPLIES	6	6	56	1019	3175	3293	3641	3646	
6361	193270	11	WATER USAGE	6	6	3351	3873	3906	4294	4676	4679	
6362	193250	12	WATER SYSTEM	6	4	3536	3925	4034	4914			
6363	193290	11	WATER VAPOR	6	5	187	2290	2431	3323	4911		
6364	193480	13	WATERFLOODING	6	6	1603	2696	2691	3227	4700	4794	
6365	193800	6	WEAPON	6	5	552	3436	3485	3406	4463		
6366	195580	14	WILL INFLUENCE	6	6	3	1979	2766	3259	3400	4821	
6367	195590	9	WILL LEAD	6	6	394	2260	2267	3004	3117	4257	
6368	202240	3	180	6	6	1117	2603	2712	3535	3707	3851	
6369	174430	23	SYNTHETIC FUELS PROGRAM	6	6	201	1428	2133	2662	2717	2893	
6370	173970	6	SWINGS	6	6	1176	1190	1213	1413	1542	4065	
6371	173990	6	SWITCH	6	5	1600	2527	3087	3911	4791		

ここで専門用語としては，エネルギーという分野を限定して考えた場合に概念が特定できるものを積極的に取り込むこととし，頻度リストを分析して最も効率的（ポスト・エディティングが少なくて済む）な専門用語候補の範囲（頻度４～８０）を選択した。

さらに次のステップとして，選択した専門用語の間で共出現状況を調べ，用語のグルーピングを行った。これはシソーラスを機械的に作成するための一つの試みであり，同一記事内に共に出現する語の間には何らかの関係があるとの考えに基づくものである。但し，シソーラス作成の完全な自動化をねらったものではなく，語の間の関連度を機械的に調べ，その結果を基に最終的な概念構造（例えば上位概念，下位概念等）の区分けは専門家の判断に委ねることとした。

(3) 今後の課題

専門用語抽出実験として，今回の方法でかなり効率的に行うことができることが確認できた。とりわけ２～４単語構成の用語抽出については相当な成果があがっている。さらに精度を上げるためには，単語の品詞情報を調べ，名詞（句）を抽出し，今回の方法に結びつけること等が考えられる。

今回はエネルギー分野のデータを使用した訳であるが，データ量がまだ不足している状況であり，今後，データ量を増やして実験を試みる必要がある。また，エネルギー以外の分野についても同様な手法（頻度分析等）を適用して結果を分析することも興味深い点である。

4. 今 後 の 課 題

4. 今 後 の 課 題

昭和５６年度より開始した文章情報の総合利用に関する調査研究事業も第３年度を終り，総合解析システムの構築に向けた種々の調査分析及び開発を通じて，多くの成果と解決すべき課題とが明らかになってきている。それらの課題には，次年度の研究開発の中で明らかにし，あるいは実現すべきものがある。また，各々の業務において，本研究が提示する総合解析システムを実際に適用し，構築するに当って実現すべき項目としてここで提示する課題がある。更に文章情報データベース総合解析システムの完成・稼動のためには各種辞書の開発をはじめとして，より深く追求していくべき，残された課題もある。ここでは各機能別に明らかになってきた課題について(1)で述べ，それに対して本調査研究において取り組むべき課題を(2)に示す。

(1) 各機能における課題

研究開発を行ってきた中で明らかになった，あるいは残された課題を総合解析システムの概念像における機能に，ほぼそって各々見ていく。

① 原情報に対して，キーワード付けや構文・意味処理などが施された結果としての二次情報データベースの内容に対して定量化分析を行っていくという種々の手法に関して調査研究を行ってきた。また，いくつかの手法に関しては実際に実験を行ってそれを検証するとともに，開発も行ってきた。

それに対しては，まず分析手法の充実に向う必要がある。とくに記事情報の内容把握を行うための分析手法の充実が重要である。

総合解析システムの中にこの分析機能を位置づけるときは，個別の分析機能の充実もさることながら，基礎解析機能とのリンクをとった内容分析機能の実現が何よりも重要となる。現在，定量化分析の為には，多くの場合，人手によるキーワード付け，概念付けを基としているが，これらの中で出来る限り多くの部分を自動化し，基礎解析機能として実現

する必要がある。そこでは基礎解析における構文・意味処理の結果を積極的に用いた分析手法も可能となり，また，要求されることとなろう。そして，基礎解析のインターフェイスとしての中間表現（二次情報データベース）の整理が必要である。これは文章情報の分析のニーズと，分析技術の可能性からの要因と，基礎解析機能における実現可能な技術と拡張可能性からの要因の双方から調整しあう中で明らかにすべきものである。

② 検索機能

データベースの作成・更新機能，データベースの検索機能の開発を進めてきた。その中で，いくつかの解決と実現が望まれる課題がある。まず，より柔軟な検索機能の実現の方向へ進む必要がある。この中で，内容検索あるいは自然文検索といった高度な検索を可能とすることもできるであろう。

また，総合システムの中における検索機能としては，内容分析機能の場合と同様に，基礎解析機能とのリンクが重要である。これによって，検索機能の高度化自体も可能となる。更に，分析機能とのインターフェイスも意識する必要がある。データベースに対する検索結果に対して，種々の分析手法を適用するとともに，それらの分析結果を検索の対象とすることも，総合解析システムの実現にあたっては考慮すべき課題であろう。更に，データベースに対する検索ニーズに有効に対応する為には辞書やシソーラスを利用することにより，めざましい効果を発揮するであろう。更に，翻訳機能の積極的な活用，とくに自然文生成技術を用いることによって，検索結果を利用しやすい形態にするといったこともできる。これらに向けた研究開発の必要性も，ここで提起すべき課題であろう。

③ 基礎解析機能

文章情報に対する各種の解析技術に関して調査・研究を進めてきた。

また，用語の自動抽出や文章の解析などについて実験を行ってきた。これらをふまえて，総合解析システムの検索機能，翻訳機能といった各要素で用いる二次情報データベースを自動的に作成する機能を開発することが肝要である。この技術は，同時に辞書，シソーラスといった知識ベースの作成機能とも密接した関係をもっており，あわせて考慮する必要があろう。

④ 翻訳機能

内外における機械翻訳技術に関する動向とその内容に関して，調査・分析を行い，そこで用いる種々の自然言語処理技術に関しても研究を進めてきた。それらを踏まえて，文章生成機能を総合解析システムの中で実現することが必要である。その為には，翻訳機能の側からも中間表現としての二次情報データベースの検討が要請される。また，翻訳機能に必要な辞書や文法に関する検討も，今後の課題として残されている。

⑤ データ整備

これまでに，エネルギー関連内外記事情報の整備・加工を進めてきた。その中で，総合解析システムの一部稼動を行うために必要なデータは相当程度揃ったと言えよう。一方，総合解析システムの実現の為には，辞書・シソーラス・文法といった知識ベースに関する整備・加工を行っていくことの重要性はより明確化してきた。そこで，これまでに行ってきたデータ整備作業の成果を整理するとともに，知識ベースの実現への方向を提起する必要がある。

(2) 総合解析システムの実現に向けての課題

前節で，各機能毎にそれぞれの現状をふまえて諸々の残された課題に関して述べてきたが，総合解析システムの実現のために，本プロジェクトにおいて更に具体化すべき課題として次のものがあげられる。

① 二次情報を自動的に作成する解析機能の開発

文章情報から，それに付された書誌的情報の他に，そこで使われてい

る用語を抽出する。あるいは，文の構造に関する解析を行い，総合解析システムで認識されるデータベース構造－中間表現へ変換する。

② 二次情報を用いた定量化分析機能の開発

基礎解析機能などにより得られた二次情報から，文章情報あるいはその特定の集まりが指し示す新しい情報を得るための統計的な手法や，文章情報の内部表現の操作手法を具体化する。

③ 自然文生成機能の開発

解析機能，分析機能などで得られた二次情報表現から，我々にとって受け容れやすい自然言語の形で出力する為の基本機能を開発する。

これら諸機能を準備することによって，本年度迄に開発済のものと合わせ，総合解析システムの姿を明らかにすることができる。

また，諸機関において総合解析システムを実現するに際して要求される総合データベースとしてのシステム要件に関しても，これらの開発・実験の中で，調査・研究の成果もあわせて明示することができよう。

このような総合的文章情報処理システムにおいては，ここで云う二次情報データベースの構造（システムの中間表現）の策定がその成否を決するといって過言ではない。本プロジェクトにおける研究開発が，この為の指針を供するものとなるべきである。

総合解析システムの実現に際しては，各種辞書，シソーラス，文法といった知識ベース機能の作成・整備が重要である。その為には，文章情報の潜在的・顕在的利用ニーズと技術的可能性をあわせた具体的な分析と多大なる開発・整備作業が必要であるが，本プロジェクトによる研究開発の経験は，その為の道筋を示すとともに，その重要性をより一層明らかにしてきた。その様な課題に対して，各方面において本格的な活動がなされることによって，文章情報を真に有効利用する総合解析システムは全面開花するであろう。

参　考　文　献

（3.2節）

1) 高橋達郎：情報検索，東洋経済新報社（１７６８）

2) 藤崎哲之助：動的計画法による漢字仮名混り文の単位切りと仮名ふり，自然言語処理，２８－５，１９８１

3) 森崎正人：新聞記事を対象としたキーワード自動抽出システムの一構成法，昭和５８年度電子通信学会組合全国大会論文集（１９８３）

4) 福島又一，上田修一，中山和彦：JAPAN／MARC書名のKWIC 索引システム，ドキメンテーション研究，３３(２)，１９８３．０２

5) 嘉手川繁三，脇田修躬：日本語点訳システム，情報処理学会第２７回全国大会講演論文集（１９８３）

6) 森崎正人，中園薫：キーワード自動抽出をねらいとしたシソーラス構成法について，情報処理学会第２７回全国大会講演論文集（１９８３）

7) 石井健一：新聞記事データからキーワードを自動抽出する試み，ドクメンテーション研究，３３（１７），１９８３．１１

（3.3節）

1) Luhn .H.P.

A Statistical Approach to Mechanical Encoding and Searching of Literary Information.

IBM.J.,309～317, 1957,Oct.

○ Luhn .H.P.

The Automatic Creation of Literature Abstracts.

IBM.J., 159～165, 1958,Apr.

2) Stiles,H.E

The Association Factor in Information Retrieval.

J.ACM., 8(2), 271～279, 1961.

3) Salton,G.
Dynamic information and Library processing.
Prentice-Hall Inc., 159p., 1975年
Chapter 3 Automatic Indexing and Abstracting
○ Salton,G., & McGill,M.J.
Introduction to Modern Information Retrieval.
McGraw-Hill Book Company, 448p., 1983年.
Chapter 3 : Text Analysis and Automatic Indexing, P.52～117

(3.5節)

1) Kowalski,R., "Predicate Logic as Programming Language," IFIP-74, pp.569～574.
2) Clocksin,W.F.and Mellish,C.S., "Programming in Prolog", Spring-Verlag,1981.
3) 横井, 渕, 「推論機構を内蔵した述語論理型言語Prolog」日経エレクトロニクス・ブックス, 人工知能, 1983.
4) Codd,E.F, "A Relational Model of Date for Large Shared Data Banks", Comm ACM, Vol.13, N0.6, pp.377～387. 1970.
5) 日本電気㈱「FFKO1 INQ概説書」リファレンス・マニュアル.
6) 日本電気㈱「Shape Up 仕様書(V1)」リファレンス・マニュアル.
7) 村木, 市山, 「機械翻訳に関する一つのアプローチ」, 情報処理学会, 自然言語処理研究会, 31-6, 1982.
8) 安川, 田中, 「Prologによる形態素処理と熟語処理について」, 情報処理学会, 自然言語処理研究会, 32-4, 1982.

昭和 59 年 3 月発行

発行所　財団法人 日本情報処理開発協会
東京都港区芝公園3丁目5番8号
機械振興会館内
Tel (434)8211(代表)

印刷所　株式会社 タケミ印刷
東京都千代田区神田司町2－16

58-S 003