

データベース構築促進及び技術開発に関する報告書

レコードマネジメント用辞書管理システムの開発研究

平成 3 年 3 月

財団法人 データベース振興センター

委託先 (株) オフィス 総研

本報告書は、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作成したものである。

序

データベースは、わが国の情報化の進展上、重要な役割を果たすものと期待されている。今後、データベースの普及により、わが国において健全な高度情報化社会の形成が期待される。さらに海外に対して提供可能なデータベースの整備は、国際的な情報化への貢献および自由な情報流通の確保の観点からも必要である。しかしながら、現在わが国で流通しているデータベースの中でわが国独自のものは3割にすぎないのが現状であり、わが国データベースサービスひいてはバランスある情報産業の健全な発展を図るためには、わが国独自のデータベースの構築およびデータベース関連技術の研究開発を強力に促進し、データベースの拡充を図る必要がある。

このような要請に応えるため、(財)データベース振興センターでは日本自転車振興会から機械工業振興資金の交付を受けて、データベースの構築および技術開発について民間企業、団体等に対して委託事業を実施している。委託事業の内容は、社会的、経済的、国際的に重要で、また地域および産業の発展の促進に寄与すると考えられているデータベースの構築とデータベース作成の効率化、流通の促進、利用の円滑化・容易化などに関係したソフトウェア技術・ハードウェア技術である。

本事業の推進に当って、当財団に学識経験者の方々に構成されるデータベース構築・技術開発促進委員会（委員長 山梨学院大学教授 蓼沼良一氏）を設置している。

この「レコードマネジメント用辞書管理システムの開発研究」は平成2年度のデータベースの構築促進および技術開発促進事業として、当財団が㈱オフィス総研に対して委託実施した課題の一つである。この成果が、データベースに興味をお持ちの方々や諸分野の皆様方のお役にたてば幸いである。

なお、平成2年度データベースの構築促進および技術開発促進事業で実施した課題は次表のとおりである。

平成3年3月

財団法人 データベース振興センター

平成2年度 データベース構築促進・技術開発委託課題一覧

分野	課題名	委託先
社 会	1 形態学的コメントを含む病理データベースのフィージビリティ調査 2 災害情報データベース支援環境の構築 3 AV/MARCのための分類索引データベース構築 4 気候情報データベースの構築 5 健康の自己管理と病気予防データベースの構築 6 シルバーエイジの実態及び生活に必要な情報のデータベース構築のための調査研究 7 交通事故調査データのデータベース化に関する調査研究	(株)エス・ピー・オー (株)防災都市計画研究所 (株)ダイソメディアサービス (株)エムテーエス雪氷研究所 (株)コンピュータコンビニエンス 美崎高齢者福祉互助会 美崎生活館 (株)日本自動車研究所
地域活性化 中小企業振興	8 アジア太平洋交流データベースの課題性の研究 9 戦略商圈レベルに細分化した地域データと分析・提案手法を統合化した企画支援システムデータベースの構築 10 ネットワーク化された地域情報データベースの有効なマネジメントについての調査研究 11 徳島市中小商業振興データベースの構築 12 九州地域の人材情報データベース構築	(株)西日本新聞社 バラシュート情報開発研究会 札幌凸版印刷(株) セントラル開発(株)情報図書館 RUKIT (株)ニューメディア徳島 (株)九州産業技術センター
海 外	13 海外向け国内先端技術分野中堅企業情報英文データベース構築 14 海外規格(ソ連邦国家規格)データベースの整備 15 政府開発援助(ODA)に関するデータベースの構築調査 16 専門用語データベースシステムの機能に関する調査研究 17 専門家データベース構築事業	エムライソインターナショナル(株) 日本電子計算(株) (株)日本国際協力システム アイ・エヌ・エス(株) (株)海外貿易開発協会
技 術	18 VAN用データベース管理システムの開発 19 レコードマネジメント用辞書管理システムの開発研究 20 建築CAD用拡張可能データベースのプロトタイプ作成 21 先進複合材料データベース・プロトタイプの作成 22 マイクロコンピュータのプログラマブル周辺デバイスのデータベース化 23 書誌データベース用ダイナミック・シソーラスの可能性調査と実験	シャープ(株) (株)オフィス総研 三菱電機(株) (株)次世代金属・複合材料研究開発協会 (株)日本システムハウス協会 (株)紀伊國屋書店

目 次

1. 研究の目的	1
2. 研究の方法	
2. 1. 辞書管理システムの開発	2
2. 2. 実験用辞書の作成	2
2. 3. 実験の方法	6
3. 辞書管理システムの現状と問題点	
3. 1. 辞書管理機能の課題	7
3. 2. 辞書管理機能の現状	9
4. 辞書管理システムの概要	
4. 1. システムの設計	57
4. 2. システムの操作	68
4. 3. モジュール構成	91
4. 4. 辞書の構成	92
5. 実験・評価	
5. 1. 実験	93
5. 2. 結果及び考察	93
6. 結論	94
調査文献一覧	95
〈参考文献〉	96

1. 研究の目的

近年、社会環境の変化と共に、オフィスのかかえる問題も大きく変わりつつある。

その背景となるものに、『情報の多様化による文書量の増加（紙の膨大化）』『統一性に欠けたOA化の進展』『オフィス部門への投資の遅れ』『個人の生活レベルの向上』があり、『ストレスの増加』『業務効率の低下』『狭くて乱雑』『家庭環境とオフィス環境のギャップの増大』等の問題が高まりつつある。

その結果、より快適で効率的なオフィスをめざして『オフィス環境改善』や『レコードマネジメントの確立』への必要性が高まっている。

レコードマネジメントとは不必要な記録を作成・保有せず、価値ある記録を経営資源として活用するため、記録の作成から処置（廃棄、永久保存）を経済的かつ効率的に行うしくみ（国際レコードマネジメント協会定義）をいう。

このような背景からオフィスでは、『レコードマネジメントの確立』への取組みと、あわせて管理媒体の関心は紙中心（ペーパーファイリングシステム）から機械可読化媒体（ワープロ用磁気媒体、光ディスク）を対象とした電子ファイリングシステムへと移りつつある。

その電子ファイリングシステムでは、文書のタイトル、作成者、分類記号、キーワード等の書誌事項検索システムの他に、自然語による文書内容検索システム(辞書¹)を実現しようとする動きが活発化している。しかしながら自然語検索（フリーターム検索）のままでは検索モレ等の問題が発生するため、同義語・類義語でマッチングができるようにするための同義語・類義語の辞書が不可欠になる。しかしその作成・更新には人手が掛り、辞書のメンテナンスを継続的に行うことは容易ではなく、その為にシステムの運用を諦める企業は少なくない。

現在の一般的な辞書の更新方法は、辞書ファイルをハードコピー上で行い、その後辞書ファイルに入力する方式がとられていて、人手による処理が多く、システム化が非常に遅れている分野である。

そこで、レコードマネジメントのための自然語検索システム用の辞書更新から人手作業を軽減するための支援システム機能を開発することを目的とする研究を行う。

2. 研究の方法

研究の方法は大別すると『辞書管理システムの開発』『実験用辞書の作成』および『確認実験』からなる。

以下にその概要について述べる。

2.1. 辞書管理システムの開発

このシステムはパーソナルコンピュータを用いて開発した。一般オフィスでの利用を考え、開発のツールは普及率も考慮して以下のような汎用性のあるものを選定した。

- ・ パーソナルコンピュータ：NEC製 PC-9800シリーズ
- ・ OS：MS-DOS
- ・ DBMSパッケージ：dBASE III

また、辞書管理システムの開発にあたって次の機能を開発課題とした。

- ①候補語の入力を一語一語入力する方法と、フロッピーディスクからの大量一括入力する方法の両方の入力を可能にする。
- ②候補語の出現回数により、辞書登録対象語を自動選択表示する。
- ③辞書更新時、登録対象語がどのカテゴリの配下に属するかシステム的に判断し、可能性のあるカテゴリを画面に優先表示する。

2.2. 実験用辞書の作成

実験用辞書としては「オフィスプランニング」業務に関する専門用語を事例として選定した。辞書の構造は カテゴリ — 用語（ディスクリプタ） — 同義語 からなるISA関係構造シソーラスとした。辞書の作成方法は「オフィスプランニング」を社内で担当しているコンサルタント、デザイナーに対し、検索キーワード調査票(表2-1)によりアンケート調査を行い、カテゴリの設定と用語の抽出を行った。（原案）

この原案をベースに社内文書資料（社内作成研修用資料、カタログ等）より用語を収集・整理し、実験用辞書（588語）とした。（表2-2,3）用語の整理をする上で留意したことは同義語と複合語の整理であり、以下にその方法について述べる。

表 2 - 1

検索キーワード調査票 (ファセット分類抽出表)

J-22-16

対 象 オフィスプランニング		セクション名 アメニティ G 氏 名	
STEP 3 : 基本計画	活用方法 (活用目的)	活用目的の 重 要 度	検索するときの キーワード
文 書 ・ 資 料 名			
基本計画書 レイアウトプラン	基本計画の作成	3 : 大 2 : 中 1 : 小	基本レイアウト名 (対向式等) 場所 (教室、ミーティング)
照明計画	基本計画の作成 商品選定	3 : 大 2 : 中 1 : 小	照明計画 照明器具メーカー名 照明方法 照明を便する場所
配線計画	〃 商品選定	3 : 大 2 : 中 1 : 小	配線計画 床材、等のメーカー名 配線方法 使用する部材
カー計画	〃	3 : 大 2 : 中 1 : 小	カー計画 カー、イヤー等 カラーコーディネート
スペーススタンダード	〃	3 : 大 2 : 中 1 : 小	配列方法 (対向式、 学校式) 一人当りの㎡ スペーススタンダード
		3 : 大 2 : 中 1 : 小	

* 3 : 重要度が高い 2 : 重要度がやや高い 1 : 重要度が低い

表 2 - 2

-----*-* シソーラス辞書一覧リスト *-*	
カテゴリー	ディスクリプター 同義語
1. 関連法規	1. 建築基準法 2. 消防法 3. 工業標準化法
2. 会社名	1. 太陽神戸三井銀行 2. 日本ガイシ 72. 日本レコードマネジメント 1. NRM 73. 関西レコードマネジメント 1. KRM 74. 東京レコードマネジメント 1. TRM 75. オフィス総研 1. オフィス総合研究所 2. ORI
3. 官公庁	1. 大蔵省 2. 通産省 3. 総務庁
4. 団体，学校名	1. 記録管理学会 2. ARMA 3. 日本事務機工業会 4. 日本機械輸出組合 1. JMEA 5. データベース振興センター 1. DPC
5. オフィス家具商品	
6. 場所・空間	1. スペース 2. エリア
7. オフィス機器	1. ワープロ 1. W/P 2. WP
8. レコードマネジメント	1. レコードマネジメント 1. 記録管理 2. レコマネ
9. 記録形態	
10. サプライ	1. ファイリングサプライ 2. バインダー
11. オフィスプランニング	1. 基本方針策定 2. コンセプト 3. 実態調査 4. 文書量調査 1. ボリューム調査 5. 近接関連分析 1. リレーションシップチャート 6. エルゴノミクス 1. 人間工学 2. 労働環境工学

表 2 - 3

* - * シソーラス 辞書 一覧リスト * - *		
カテゴリー	ディスクリプター	同義語
12. オフィス環境	1. 効率性 2. 快適性 34. 熱 35. 温度 36. 音 37. 暗騒音 38. 空調音 39. 機械音	1. アメニティ 1. 吸房 2. 音環境 1. 背景音
13. 一般用語	65. 国外 66. 国内 67. 作成 68. 策定 69. 仕事 70. 支援 71. 支店 72. 施工 73. 自社 74. 執務 75. 社屋 76. 従業員 77. 重	

2.2.1 同義語の整理方法

同義語はつぎの 8 つの観点から整理した。

- ① 意味的：デスク；机
- ② 略語：NRM；日本レコードマネジメント
- ③ 表記の相違：ITO；アイティーオー
- ④ 送りがなの相違：読み込み；読込み
- ⑤ 異字体表記：製鉄；製鐵
- ⑥ アクロニウム：CI；C. I.
- ⑦ 長音記号付与の有無の相違：レコードマネジメント；
レコードマネージメント
- ⑧ 外来語表記：硝子；ガラス

2.2.2複合語の整理方法

複合語については可能な限り、分割した。ただし、検索時のノイズをできるだけ少なくするためにも次の2つの場合は分割せずにそのまま登録した。(参照2)

- (1) 分割すると一方の文字が一文字になる場合

(例) 分類表→×分類:表

- (2) 分割後の単語がどちらとも、元の複合語と同様に、または同時に発生する可能性が高いと思われる場合

(例) 配線設備→×配線:設備

2.3.実験の方法

サンプル文書から単語を抽出し、その抽出したものを辞書管理システムに入力・更新し、今回開発したシステムの機能を検証するための確認実験を行った。

以下にその手順を示す。

- (1) 単語の抽出

サンプル文書(企画書)から単語を抽出し、テストデータ用のFD(フロッピーディスク)にファイルした。

- (2) 辞書管理システムへの入力

テストデータ用FDから辞書システムに一括入力した。

- (3) 更新確認実験

システムを起動させながら、目標とする機能を確認した。

3. 辞書管理システムの現状と問題点

情報検索システムに用いられるシソーラスを中心とする辞書管理機能について考察し、現状を文献を基にレビューし、当委託課題の参考にする。

3.1. 辞書管理機能の課題

ある一つの分野（多くは専門領域、例：学問分野）において情報交換および情報流通を行う上で、専門的概念（カテゴリという）を表現するために、その表現形式を定め、その下に設定された単語の使用が必要になる。この様な単語を用語もしくはディスクリプタという。日常のコミュニケーションにおいて用いられる単語は単に単語もしくは語という。

単語には、同一概念を表現する複数の異表記語が存在する。また、人間は、一つの概念を表現するのに広範囲のレベルあるいは狭義のレベルから考え単語表現をする。そこで、ディスクリプタを中心に単語の関係を明示するためにシソーラスが構築、利用されてきた。

データベースの構築における索引データの付与および情報検索における検索データの設定にシソーラスが用いられている。索引データおよび検索データは、情報内容および検索要求内容をシソーラス用語によって表現される。

シソーラスは、用語間の関係をディスクリプタを中心に展開した構造化用語集である。その構造はディスクリプタを中心に上位概念語、下位概念語、同義関連語を展開し、さらに各語と他ディスクリプタとの関係をも表現したものとなる。

専門領域の発展にともない、概念の変遷、新規用語の出現等が生じる。それに伴う用語間の関連づけの変更、新規用語の登録等が生じる。索引データの付与および検索データの設定のために常にシソーラスをメンテナンスする必要がある。ここに、シソーラス管理機能の課題が生じる。

また、索引データの付与および索引データの設定のためのディスクリプタの選択に対し、索引作成者および情報検索者の情報内容および検索内容をディスクリプタに接近させるためのシソーラス用語検索機能の課題が生じる。

さらに付与された索引データと設定された検索データとの間の、検索処理時における照

合機能の課題もある。

上記のいずれの課題も、その中心をなす技術的問題は、索引作成者および情報検索者の情報内容および検索内容を表現する用語とシソーラス内にエントリされている各用語との照合機能の実現に尽きる。

以下、具体的に例を上げて示す。

3.1.1.形態一致機能の課題

①同義異表記語データは照合できず欠落が生じる。

例：〔電子計算機〕に対して〔コンピュータ〕は照合できない。

②短縮型表記語データは照合できず欠落が生じる。

例：「電子計算機」に対して「電算機」は照合できない。

この結果、類義語データは照合できず欠落が生じる。

この①の問題解決には、シソーラスに全ての同義異表記語を登録する以外にない。

②の問題解決には、短縮型表記語生成機能の開発が必要になる。しかし、全ての生成規則を発見することは容易なことではない。そこで、①の問題解決と同じくシソーラスに全ての同義異表記語を登録する以外にない。いずれの課題も自然言語処理機能のレベルでの研究が必要であり、現時点において解決されていない。

そこで、現実には、データベースの構築において、シソーラス用語以外に自然語索引データを設定し、索引データの網羅性を補償している。一方、検索渡れを防ぐために、索引機能としてトランケーション機能を実現している。

3.1.2.意味的一致機能の検索

①異表記複合語データは照合できず欠落が生じる。

例：「図書館業務システム」に対して「図書館業務のシステム」は照合できない。

この問題解決の為の研究は、上記1)の問題解決同様、自然言語処理機能の課題でもあり、現時点において解決されていない。

そこで、現実には、データベースの構築において、上記の例においては、例えば、「図書館」、「業務」、「システム」と分割し索引データを設定し、索引指示機能（論理式展開）委ね、検索を可能にしている。ただし、検索率は補償されるが検索精度は落ちる。

以上の問題解決に、現行のデータベース構築において、シソーラスの補遺版の作成、検索法マニュアル等を作成し、シソーラスを利用する上での対処策を計り実施している。しかし、これら方式は、予防策であり根本的な問題解決になっていない。

3.2. 辞書管理機能の現状

レコードマネジメントを目的に使用する索引語のファイル（＝辞書）に対し、新規の索引候補語を更新するシステム機能が必要になる。その機能は、情報検索システムに用いられるシソーラスの課題と同じである。そこで、辞書ファイルに対し直接更新を行う支援システム機能を開発する為に、その課題解決策について、以下研究開発の現状を文献調査を元にレビューする。

- (1) シソーラスの持つ一般的な問題点、シソーラスの構築課題およびメンテナンスを行う上での技術的課題を、わが国を代表するシソーラスの一つであるJICSTの「JICST科学技術用語シソーラス」について文献11)、7)、5)、および新聞記事データベースの構築・検索システムに用いられている「NK-MEDIAシソーラス」について文献12)を分析する。文献11)では「JICST科学技術用語シソーラス」の開発経緯を、文献7)では、1987年に行われた改訂作業内容を、文献5)ではその後の課題を、文献12)では「NK-MEDIAシソーラス」の開発経緯およびその後行われた第1回改訂作業について報告している。
- (2) シソーラス内のカテゴリのレベルは、索引データの付与および検索データの設定にディスクリプタを発見するために、また新規索引候補語の更新の際に、特にその概念範囲の理解に、大きく影響する。各シソーラスのカテゴリ・レベルについて文献9)で実際の複数のシソーラスを比較評価している。カテゴリの変遷に対するメンテナンスについて文献1)および10)で、また概念間関係を明示する一つの手法としてJICSTにおいて実現し利用に供している“タームチャート”が文献6)に示されている。

- (3) 新規索引候補語を更新する技術的課題について、文献13)では上記3.1.で示した異表記語の管理問題について整理し、現実的な辞書管理システムを示している。新規索引候補語の更新について、文献3)ではカテゴリ間の関係において行う問題を、文献8)では用語間関数を用いての方法を、文献4)では文節レベルでの機能語情報を用いる方法を、文献14)では日本語の語基レベルでの形態的一致について大容量の用語を基に分析した結果を示している。
- (4) シソーラスのメンテナンスを実際に実現するためのシステム機能およびその効率等の課題は、対象とするコンピュータのレベルと合わせ重要な課題である。文献2)はミニコンピュータのレベルのシステムを紹介している。
- 以下に、各内容の概要を示す。(文末に文献一覧を掲載)

3.2.1. 調査文献概要

- (1)文献1) M.F.ブルンデット(M.F. Brundet): 情報検索システムにおける自動的シソーラス更新の概念設計; A CONCEPTUAL FRAMEWORK FOR AUTOMATIC AND DYNAMIC THESAURUS UPDATING IN INFORMATION RETRIEVAL SYSTEMS, Colling80, pp.586-589(1980)

冒頭にABSTRACTとしてまとめられているように、書類検索や新分野開設を目的とした自動シソーラス構造の方法論を呈したのが本論文である。シソーラスに含まれる情報が、Andやorといった自動的なユーザーのフィードバックを経て新たな検索の方式を導き出す。理論上ではあるが、データベース構成へ向けての提示である。

シソーラスは通常“ディスクリプタの集合体であり、ディスクリプタ間の関係情報の集合体である。”と定義されている。データベースに新しい文書を入力するとき、キーワード作成のため、文章からデータを引き出し系統化する。まず文法的変格が定型に直されるのが第一義だろうがこの直された型をitemと呼び、本論文でシソーラス構成を論ずる上で基本単位となっている。

まず、局所的なことから論じている。R.アター氏(R. Attar)とA.S.フランケル氏(A. S. Fraenkel)の共同論文に習い、局部の群生化する。この時、対象となる文書群はD1、1セットのitem群はT1、文書番号はDN、文章番号はSN、同一文章内のitem番号はIN、item内の単語個々はその番号をiやjでwt(i)のように示し、i番めのtとj番めのsの間隔

を d 、 d の機能を F 、関係する s 、 t の単語の合計値を b 、標準化させる機能を $\mu R(s, t)$ 、参照するベクトルを R 、仕事としてのベクトルを $E R$ とし、その関係を計算式で求めた。

$$\text{式 ① } d = |IN_t(i) - IN_s(j)| \quad \text{with} \quad \begin{cases} DN_t(i) = DN_s(j) \\ SN_t(i) = SN_s(j) \end{cases}$$

式 ② Let be F a function of the distance d :

$$F(w_t(i), w_s(j)) = \begin{cases} \cdot 1/d \text{ if } w_t(i), w_s(j) \text{ are} \\ \text{in the same sentence} \\ \text{with } d \leq 20 \\ \cdot 0 \text{ otherwise} \end{cases}$$

式 ③ For a and $t \in T$ we define :

$$b(s, t) = \sum_{T} \sum_{T} F(w_t(i), w_s(j))$$

さらに新情報によるシソーラス更新の考え方として、新たな用語と概存するitem間の類似性や近接性を検討する。このとき M 、 m といったitem s 、 t 間の関係を示す機能2つを定義し(式④、⑤)、概存の参照するベクトルを R_s 、新しいものを R' とし、その標準化のベクトルとの連合機能の重要性を見い出している。(表3-1)

$$\text{式 ④ } m(s, t) = \frac{\text{Min}(\mu_{R_s}(s, t), \mu_{R'_s}(s, t))}{1 - |\mu_{R_s}(s, t) - \mu_{R'_s}(s, t)|}$$

$$\text{式 ⑤ } M(s, t) = \frac{\text{Max}(\mu_{R_s}(s, t), \mu_{R'_s}(s, t))}{1 + |\mu_{R_s}(s, t) - \mu_{R'_s}(s, t)|}$$

次にシソーラス構築、更新における m 、 M の機能の役割に考究する。ここでitemの x を用い、item群 T に属する x に関し、 R_x 、 PS_x 、 TX の3つの機能ベクトルを定義する。

$$PS_x = \{t / \mu_{ps}(x, t) > \alpha\}$$

x に関するitemは概存シソーラス内の1文脈、或いはいくつかの文脈、或いはもっと全体的な文脈の内にある。3つのベクトルは0と示すこともあり、その場合によって対処が

表 3-1 Using the above functions m and M (formulas (4),(5))

$\mu_{R_S}(s, t)$	$\mu_{R'_S}(s, t)$	$m = \frac{\text{Min}(\mu_{R_S}, \mu_{R'_S})}{1 - \mu_{R_S} - \mu_{R'_S} }$	$M = \frac{\text{Max}(\mu_{R_S}, \mu_{R'_S})}{1 + \mu_{R_S} - \mu_{R'_S} }$
0	1	indeterminate	0.5
0	0.2	0	0.166
0	0.8	0	0.44
0.1	0.9	0.5	0.5
0.1	0.8	0.33	0.47
0.1	0.7	0.25	0.43
0.1	0.6	0.2	0.40
0.1	0.5	0.16	0.36
0.1	0.4	0.142	0.30
0.1	0.3	0.125	0.25
0.1	0.2	0.11	0.18
0.1	0.1	0.1	0.1
0.5	0.5	0.5	0.5
0.9	0.2	0.66	0.52
0.9	0.4	0.8	0.6
0.9	0.5	0.83	0.64
0.9	0.6	0.85	0.69
0.9	0.8	0.88	0.81

異なる。まず1番めはPSxもTxも0ではないとき——更新に3過程があり、その第1は、新データの処理としてRxとPSxを計算し、比較する。第2に更新の手順として2とおりがある。仮の参照ベクトルPxとTxが同一itemを有するならETxはmにより修正し、有しないならTxはMにより修正する。第3に仮シソーラス更新の手順としてPSxの連合機能を μ_{Rx} , μ_{PSx} , Mを使って計算する。対処の2番めはPSxとTxが0であるとき——この場合、どの文脈にもxはない。対処の3番目はPSxは0だがTxは0でないとき——mを

使用し、 $R_x \neq \phi$ なら T_x をシソーラスに築く。更新を繰り返すと付加的な item よりも全体的な item 価値が、シソーラス更新に伴って増大していくことが計算できる。

ここで、シソーラスの全体的処理として、一組の参照ベクトル T_x 、 T_y について類似性 $r(x, y)$ 、間隔 $d(x, y)$ を測定する。(式⑥, ⑦)

$$\text{式 ⑥ } r(x, y) = \frac{\sum \text{Min}(\mu_{T_x}, \mu_{T_y})}{\sum \text{Max}(\mu_{T_x}, \mu_{T_y})}$$

$$\text{式 ⑦ } d(x, y) = 1 - r(x, y)$$

is a pseudo-distance whose range is [0,1].

この T_x 、 T_y の示す item 群 T はより大規模ではあるが、6000 語を超える様な超大な辞書には現実的ではない。この点の克服のために局部的な組みあわせ行列から全体的なものを築くことが考えられる。

そして更新という作業にはフィードバックが重要になってくるが、このことに関しては様々な論文が書かれている。しかしそれらはある J. カーティン氏(J. Courtin)によって一論文にまとめられ、それは局部的場合にも、全体的場合にも、局部的・全体的あわせた場合にも応用できる。最も良い方法としてはその局部的・全体的双方の文脈に使用できることのようなが、立証はまだされてないのである。

本論文に於て、考察の基にしたシソーラスはおよそ平面的であり、今回は集められた内容を反映した item かと、優位性とかは予想していない。その上で、今後のシソーラスに関する意見を受け入れ、関連する関数も一部値を変化させてゆき、類義語の定義は豊富にしてゆくことを予想している。

次段階の目標として以下の 2 点をあげられる。

① キーワードがあやふやでも使用でき、不足のない質問文を作成すること。

② 自動索引のメカニズムを設置し、文書の集団化(分類化)を図ること。

こうした点でシソーラスが構想されるべきである。

- (2) 文献2) E. J. カザラウスカス(E. J. Kazlauskas), T. D. ホルト(T. D. Holt): シソーラス作成へのミニコンピュータの利用; The Application of a Mini computer to Thesaurus Construction, Journal of the American Society for Information Science vol.31 No.5, pp.363-368(1980)

N I C S E M (National Information Center for Specical Education Materials)用データベースのためのシソーラスを作成するいろいろなフェーズでミニコンピュータを活用したことについて述べる。

初期値の用語の選択と整理は下請けであるV O R Tコーポレーションが、N I C S E Mの情報科学および専門教育のスタッフと共にを行った。

N I C S E Mが必要とする用語はシソーラスで管理されるべきものであった。しかし、残念なことに、このシソーラス作成に利用されるはずのこの専門分野においては共通の専門用語、あるいは体系は存在しなかった。

専門教育の学生のニーズに合わせようとした多くの分類法や管理システムが開発されたが、これはという均一の評価を得られるものはなかった。しかしながらこれらの分類法や管理システムはシソーラス用語作成に活用されるべきである。いくつもの異なったソースから用語を体系づけて手作業で整理するには、大量文書の作成やタイプ、編集、並べ換えなど、事務的な時間が必要になる。これらの仕事はまさにコンピュータに向いていると思われた。

(a) コンピュータの特徴と開発のプロセス

このタイプのコンピュータ処理に求められる性能は、オンラインでの用語の入力、編集(並べ換え・グルーピング)、用語データベース作成のためのテキストのハイスピード分析、個々のシソーラスインデックスフォーマット出力である。

V O R Tコーポレーションは、ヒューレットパッカード(H P) 1000ミニコンピュータを使った。H P 1000はハイスピードプロセッサ(595nsec), マルチタスクオペレーションシステム(R T E), 中位のオンラインディスク容量(15MB)を有する。言語はF O R T R A Nを使用した。オンライン上で新用語の加工・追加をし、インデックスを変更した用語をマージするのに有効な直接アクセスを用い、更新作業には十分なスピードが得られる。

他のシソーラスについてレビューした後、N I C S E Mのシソーラス構築のための開発

方針を決定した。

この方針の重要な点は専門的分野、用語の編集、専門教育への幅広い研究などを考慮することである。分野、用語の編集、専門教育への幅広い研究など全体を通してみると、次の5つの基本的なセグメントに分かれる。(図3-1)

それは、①用語の収集、②用語の編集／合成、③用語のグルーピング、④用語の体系化、⑤インデックスのプリントである。

HP1000によるオンラインでのシソーラス作成はセグメントの②～⑤により展開した。

(b) 用語の編集／合成

用語の収集プロセスにおいて、初期に16,000語を集めた。これらの用語は専門教育に参考になる可能性のあるものをベースに収集した。

認知された用語はワークシートに記録するか、あるいは原文に直接マーキングした。

入力オペレーションはステップごとに完全コンピュータのCRTに指示がしている。オペレーターがCRT上で用語を入力する指示や、“新用語”のタイトル選定を行なう。システムがオペレーターに対して必要な用語のデータ(題目、ソースナンバー、スコープノート)について聞いてくる。オペレーターは1時間あたり平均250用語以上の入力処理ができた。

用語ファイルにあるすべての用語はソートされ、アルファベット順にプリントされた。

このアルファベット順のリストは同義語の判別、エラー修正、用語形態の確認をするために編集スタッフが利用した。

同義語は用語ファイルから物理的な削除はされずに、フラグがたてられた。

セグメント③(用語のグルーピングのプロセス)では編集スタッフによりグループの名前づけがなされた。

用語が編集され、フラグ処理をされた後、16,000語の入力語に対し、13,000語が登録候補(統制語)として残った。

(c) 用語のグルーピング

シソーラスは専門教育の学生のための専門スキルと関連づけた用語グループが求められた。

用語のなかには、プログラムの運営、アカデミックな研究または専門教育のカテゴリのタイプと関連するものもいくつかあり、用語のグループはスキル指向とノンスキルのカテゴリに区分された。いろいろな用語グループに名前が提案・選定された後、オンライ

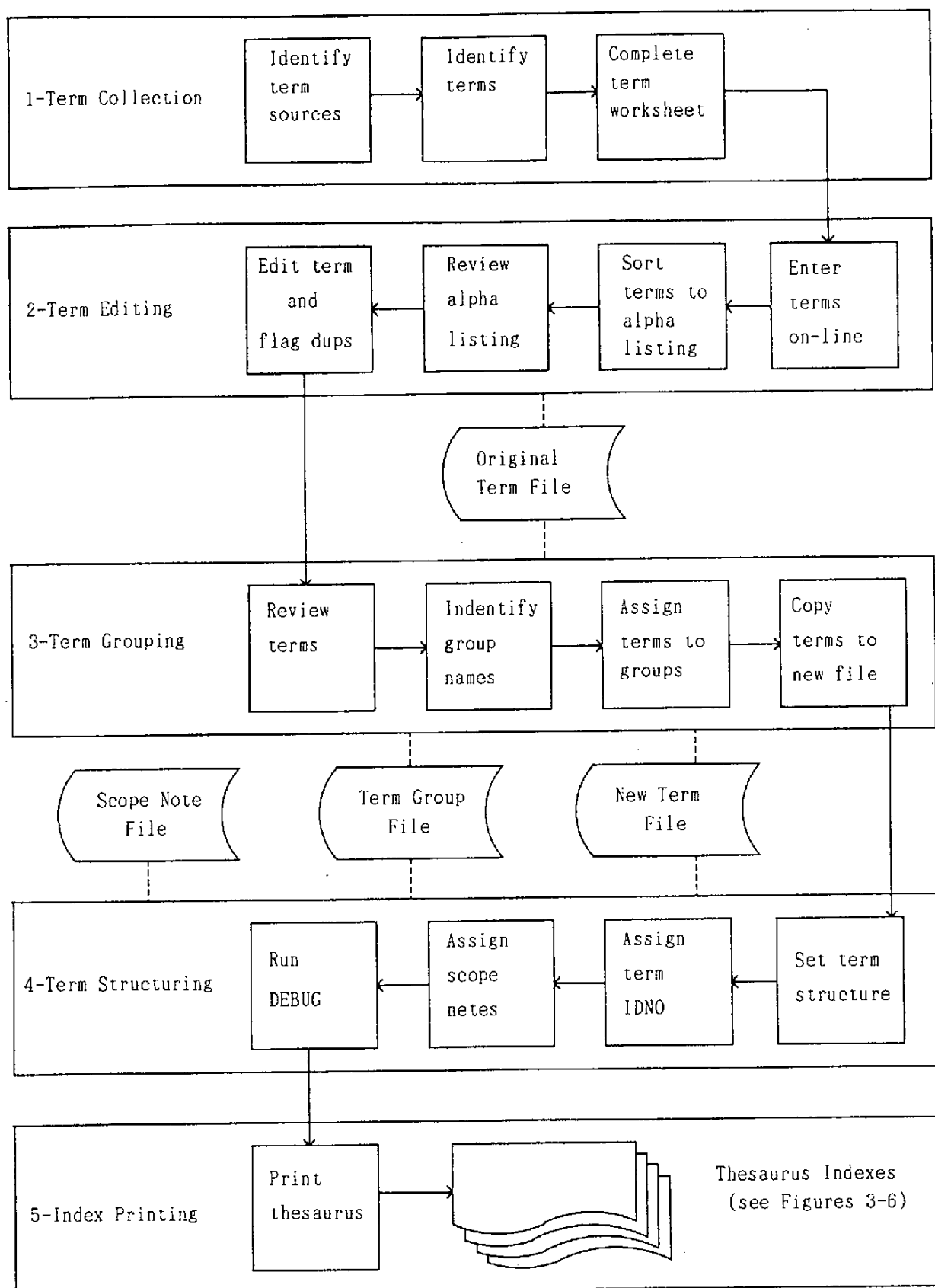


図 3-1 NICSEM シソーラス開発処理

ングループファイルが作成された。このファイルは各々のグループ名と関連しているオリジナル用語ファイル内の用語No.が含まれている。

全ての有効な用語は150グループに割り当てられたが、修正後100グループになり、オリジナルの16,000語の内5,000語のみが残され、11,000語はフラグ処理がされたものである。

さらに、5,000語の有効な用語は再検討が続けられ、2,700の利用可能な用語が残った。

(d) 用語の構造・インデックスのプリント

この2,700語は再調査され、5桁の文字数式でIDNo. (identification number)が、各用語にオンラインで付与された。

デバッグプログラムは容認可能な用語 (Use For : U F) 間の相互の関係および、容認しない用語 (Use Refernce : U R) 間の相互関係を検証するために用語ファイルに対して実行された。その結果、2060 U Fと560 U Rの用語になった。(表3-2) ~ (表3-5) に4種類のインデックスを示した。

<結論>

N I C S E Mシソーラスの作成において、用語の入力、編集、構成処理は、ミニコンピュータを利用することにより、非常に品質が高められた。特に、用語が収集され、このシステムにすみやかに入力できる点に明確に現われた。

表 3 - 2
N I C S E M
(順列索引)

Checking accounts
Savings Accounts
Accreditation
Achievement Tests
Career Acquisition
Acting
Action Verbs
Reflective Or Active Lestening
Active Passive Voice
Enrichment Activities
Leisure Activities
Playground Activities
Recreational Activities
Self Actualization
Visual Acity
Adaptive Behavior Scales
Adaptive Communication Devices
Adaptive Eating Devices
Adaptive Travel
Adding Machine Use
Addition
Addition
Addition Facts
Additon Regrouping
Column Addition

表 3 - 3

N I C S E M
(アルファベット順索引)

XN-03	Abandoned Children
HJ-0924	Abbreviations
XR-03	Ability Tests
XL-03	Abnormal Psychology
NJ-0303	Abolition
FX-0609.03	Abortions
	SN : Sex education
ZZ-1050	Abstract Reasoning
	Use : Giftedness
YJ-0303	Academic Degrees
VF-02	Academic EducationA
ZZ-2365	Academic Giftedness
	Use : Giftedness
HF-1512.15	Academic Rank
FJ-24	SN : Professional
	Acaculia
	Use : Arithmetic Disorders
	Accents
ZZ-0510	Accessibility Skills
	SN : Ability to recognize and remove or overcome an obstacle which may prevent ones entrance to or use of a building or piece o equipment
ZZ-0510	Accident Prevention
	Use : Safety
ZZ-0415	Accident Procedures

表 3 - 4

N I C S E M
(分類索引)

B	Learner Characteristics
BB-03	Exceptionalities and Disorders
BB-06	Arithmetic Disorders
BB-09	Behavior Disorders
BB-0903	Birth Problems
BB-0906	Premature Birth
BB-12	Communication Disorders
BB-1203	Agnosia
BB-1206	Anomia
BB-1209	aphasia
BB-1209.03	Receptive aphasia
BB-1209.06	Expressive Aphasia
BB-1212	Auditory Processing Disorders
BB-1215	Dysphasia
BB-1218	Dysphasia
BB-1218.03	Ezpressive Aphasia
BB-1221	Expressive Dysphasia
BB-1224	Language Disorders
BB-1227	Memory disorders
BB-1230	Receptive Language disord
BB-15	Deaf Bling Individuals
BB-18	Disadvantaged Individuals
BB-21	Emotional Disturbances
BB-2103	Autism
BB-2106	Neurosts
BB-2109	Psychosis
BB-2112	Schizophernia
BB-21115	Derious Emotional Disturb
BB-24	Giftedness

表 3 - 5 N I C S E M (分類索引チャートフォーマット)

Exceptionalities and Disorders ...	Arithmetic Disorders
	Behavior Disorders
	Birth Problems Birth Defects
	Premature Birth
	Communication Disorders Agnosia
	Anomia
	Aphasia Receptive Aphasia
	Expressive Aphasia
	Auditory Processing Disorders
	Dysnomia
	Dysphasia Expressive Dysphasia
	Expressive Language Disorders
	Memory Disorders
	Receptive Language Disorders
	Deaf Blind Individuals
	Disadvantaged Individuals
	Emotional Disturbances Autism
	Neurosis
	Psychosis
	Schizophrenia
	Serious Emotional Disturbances
	Giftedness

すなわち、編集スタッフがオンラインにより用語データに直接アクセスできることにあり、このコンピュータシステムのスピードは編集スタッフの判断と処理の時間を30%は減少したと思われる。

N I C S E Mのシソーラス作成の経験からミニコンピュータが手作業やまたは大がかりなバッチ処理システムは利用できないシソーラス構築に対して、経済的でかつ、反復できる処理が得られることがわかった。

- (3)文献3)ジェーン・エイチソン (Jean Aitchison): 社会科学におけるシソーラスの統合;
Integration of thesauri in the social science, International Classification Vol.
8 No.2, pp.75-85(1981)

本論文は、社会科学の統合されたシソーラスの特性とその編成を審議しようというものである。その試みは、書誌学に於て全分野に渡る普遍的な分類が組織されている様に、シソーラスに於ての作成である。本来シソーラスはそれぞれの専門分野で充実させるもので、統合されたシソーラスがそうした本来のシソーラスに代替するものではないが、データベース検索、参照資料として有効だろうとしている。

統合されたシソーラスの特質として次の点をあげている。主題範囲はここでは社会化学に限定している。社会科学の中には教育・心理・宗教といった専門分野があるが、社会科学というある程度グローバルな範囲にしている。1つの情報言語に対し、関係のある情報言語がリンクされているが、そのリンクされている言語のレベルを広義・狭義・関連語という様にとり、それぞれ<・>・-の符号を用いて示している。(表3-6) 言語そのものが持つ特異性にも注目し、安易に他の言語に置き換えたりせず、その語の使用される地域的広がりも分類システムとして取り入れようとしている。

表 3 - 6

=	Foreign language equivalent(s) (Definition or scope note)
=	Non-preferred synonym or quasi-synonym
<	Broader term in the same part of the display
>	Narrower term in the same part of the display
-	Related term in the same part of the display
*<	Broader term in another part of the display. Notation
*>	Narrower term in another part of the display. Notation
*-	Related term in another part of the display. Notation

Examples:

Unemployment TLE
 < Employment
 > Cyclical unemployment
 > Frictional unemployment
 > Structural unemployment
 > Temporary unemployment
 > Youth unemployment
 *< Deprivation QGP
 *-> Deflation TEJ.F
 *-> Economic recession TDO
 *-> Redundancy TUF.XK
 *-> Unemployed People QGX
 *-> Unemployment insurance QFG.G

Technical unemployment TLE.PV
 (Unemployment due to displacement of certain occupations by technology)
 = Technological unemployment
 < Structural unemployment
 *-> Technological change UAH

サイズは、15000～30000語ぐらいをみているが、これは特異的情報言語が加えられていくことによって増加していく。その特異性を示すものとして分類毎の表示方法や多種言語によって記録することを重視している。

また、辞書等といった一般の参照書によつてのチェックも必要であると示唆する。

次にディスクリプタバンクについても、スコープノートや外国語の同義語情報や用語の相互関係を付加していくものと期待している。特に見出し語について、その見出しに付随する情報言語の見出し語とのわずかな違いも見分けられるようにする必要がある。

分類は勿論重要な役割を負っている。その文書の専門分野、全景がおおまかではあるが示されている。本論文ではLC（アメリカ合衆国議会図書館分類）やUDC（国際十進分類法）をもとにするが、他の分類法ともあわせ、改善していく必要性も求めている。

それには主題の分析が必要でもあり、また分類的に主題をどうとらえるかを示さねばならない。その上で主眼となる用語と情報言語の関係をマトリックス的に捉えて示すことは理解に役立ち、主分類がどれにあたるか、機械的に解せることになる。

そこでマトリックスで用語を捉えた場合の統合シソーラスについてその処理の方法を論じている。新用語の追加から用語の削除・類義語や準類義語の付加・分類体系のレベルを変更する方法・体系（分野）の追加・関連用語の追加・合成用語の処理まで述べている。

現在は適合性を示すにはマトリックスよりも構成要素としての情報言語の持つ関連性のデータの方が、ディスクリプタ内に於てはより効果的であろうが、次第にそれも克服されていくだろう。

以上のように個々の構成要素を検討したうえで、シソーラスの設計を行う。分類は主にユネスコのシソーラスと同じようにし、ROOTのスコープノートの部分は応用した。表示法はコンピュータ検索がよりしやすい様に調整が幾分か必要ではあるが、統合されたシソーラスに最も近いBC2のものを採用した。マトリックスの紙上の表示も試みた。（表3-7）用語自体の配列はアルファベット順とした。機械処理として最も有効であるためだが、分類の包括的概観を捉えるには難しい。その他、前出した見出しとなる用語を基準にして、付随する言語が広義が狭義かの別も表示した。

また、統合シソーラスとして英語以外の言語バージョンも創られるだろう。基本図面は英語版になるだろうが、フランス語、ドイツ語又はその他の言語に改造することはできる。

こうして構成されるシソーラスはテストと更新を繰り返してより一般的な、均整のとれ

表 3 - 7

INTEGRATED THESAURUS - CLASSIFIED DISPLAY AND MATRIX

Bliss (002) Master (amended)

Matrix of ILs A - F

T	ECONOMICS AID MANAGEMENT (Cont'd)
TK	ECONOMIC RESOURCES (Cont'd)
TL	LASOUR ECONOMICS (Cont'd)
TLD.M	Employment
	
TLE	Unemployment
Add BT from E	+< Deprlvation OGP
	02.04.01
	Social problems
	+> Redundancy TUF.XK
Add RT from C	+< Daflation TES.F
Add RT from A	+< Economic recession TDC
Add RT from A,B,C	+< Unemployment Insurance DFG.G
	03.02.04
	02.03.02
	326.0309
	Unemp. Ins. policy
Add RT from A,B,D and E	+< Unemployed people QGX
	
	
	(By from of unemployment)
Delete. Treat at TUF. XK	Redundancy
Delete	Regional unemployment
TLE.NF	Frictional unemployment
	
	Retain BG2 scope note and print out In
	alphabetical section
TLE.NH Retain for future use	Temporary unemployment
TLE.QP	Cyclical unemployment
TLE.PR Make MT as A	Seasonal unemployment
	13.01.03
	13.01.03
	6243
Add RT from E	+< Seasonal employment TLE.SF
Add RT from A	+< Seasonal workers TLV.SU
TLE.PT	Structural unemployment
	13.09.02
	13.01.03
	S06.0656
	6744
Consider RTs from B	
and C here	
TLE.PV Retain Indenting although	Technical unemployment
not NT In A,B and C	
	13.01.03
	S06.0657
	Technological
	Unemployment
Add RT from A	+< Technological change UAH
	12.06.00
	(By age group)
TLE.QY Add new term from A,D and E	Youth unemployment
	13.01.03
Add	
RT from E	+< Unemployed young people QGX.LR
RT from A	+< Youth QLR
RT from D,E	+< Youth employment TLE.RL
	14.02.02

A	B	C	D	E	F
13.01.03	S06.0655	7335	Q55.40	HG	331.6
02.04.01	—	—	R71	LPM	—
Social problems	—	—	Social disadvantage	—	—
—	—	—	—	—	—
—	—	1638	—	—	—
03.02.04	—	—	—	—	—
02.03.02	326.0309	7336	—	—	368.44
19.01.02	Unemp. Ins. policy	—	R84.15	HG6	Unemp. Insurance
Unemployed	S92.0127	—	Unemployed	—	—
	Unemployed person				
—	—	—	—	HGCF	<331.6.063
					Unemployment types
—	—	—	—	—	—
13.01.03	—	—	—	HGCP	—
13.01.03	—	6243	—	HGCU	<331.6.063
—	—	—	—	HFT	Unemployment types
13.09.02	—	—	—	—	—
13.01.03	S06.0656	6744	—	HGCS	<331.6.063
					Unemployment types
13.01.03	S06.0657	6949	—	→	<331.6.063
	Technological			Structural unemployment	Unemployment types
	Unemployment			HGCS	—
12.06.00	—	—	—	—	—
13.01.03	—	—	Q55.60.20	HI	<331.6.063
—	—	—	—	HIB	Unemployment types
14.02.02	—	—	—	—	—
	—	—	Q55.60	HH	—

たものになり、オンラインシステムに反映されていくのである。

最後に問題になるのは管理面であるが、本論文ではそこまで言及はしていない。

このように、統合されたシソーラスは参照の基準として情報をより正確にという貴重な役割を有するだろう。既存する情報言語の質的向上をうながし、より包括的理解につながるものである。

本論文では最後に、論述の手腕と情報の蓄積と社会科学の専門知識がもっと十分であれば読者が理解しやすかったろうとしている。社会科学の統合されたシソーラスとしているが、例として用いている様であり、他の分野に転じても通じるものである。マトリックスの考え方がもう少し具体的に示されれば理解しやすかったかとも思う。また、本文にもあるように実際オンラインにのせた後の管理について言及はなく、その管理やシソーラスのオンライン更新のタイミングも検討する必要があるだろう。

- (4) 文献4) 山階正樹, 小橋史彦, 萩野輝雄: I D I O M方式の辞書構成法, 電気通信研究所研究実用化報告 Vol.30 No.11, pp.2767-2777(1981)

日本語情報処理システムにおける漢字入力システムの1つとして筆者達はI D I O M方式を提案している。その辞書を分類したのが図1である。その中で重要な一般語辞書と固有名詞についての論述である。

まず、収録語彙であるが、種々の資料(1)から一般文に使用される語彙数を約3万語と推定する。一般語辞書の基幹には32600語の「基本語彙」を収めた「分類語彙表」をおき、他資料を参照し、収録語数は24436語とした。カバー率は97~99%とみている。固有名詞については、姓名として姓10363語、名9696語を収録。地名として都道府県市町村レベルの3947語、企業名として株式市場1部、2部上場の企業名1711語を収録した。これらは使用頻度、知名度の高い名を選んだもので、固有名詞は使用者によって多様であることもあり、高いカバー率は求めている。私用辞書による充実が実際にも有効である。

次に同音語解析のために収録した単語の属性情報が必要となる。考えられる属性情報として、漢字属性情報(I D I O M記号)、文法情報(品詞、活用型、活用行)、関連語情報(慣用句、慣用表現)、シソーラス情報(意味分類記号)、使用頻度情報があるが、これらのうち、関連語情報、シソーラス情報は現状では問題が大きく除外し、I D I O M用辞書では、かな見出し、かな語幹、漢字表記、I D I O M記号、文法情報、使用頻度を用

いている。かな見出しは、複合語については容量の問題から対象外とし、地名についても行政区画名を含めないこととした。企業名は単語単位に分割、重複を省いた。かな語幹とは用言などで漢字表記をした場合の送りかなにあたる部分で活用語尾にならない部分のことである。（例；「美しい」の「し」、「美しさ」の「しさ」など）

こうした属性情報の持つ効果は、同音語をどれだけ一意に決められるか、同音語の数をどれだけ減少できるかで明らかにできる。そのために、同音語分布、優先順位とカバー率の関係を調査している。

同音語分布は縦軸を累積同音語数、横軸を同音語構成語数を設定した図表で示した。（図3-2-6）まず一般語辞書（図3-2）をみると、かな見出しだけの場合に比べてID IOM記号を併用すると最大同音語総数が約1/2に減少し、同音語総数も65%の割合に絞られる。文法情報とかな語幹も併用していくと、同音語同定効果は大幅に向上する。これら情報による処理として用言においては文法処理、体言においてはID IOM記号がより有効であるという相補性を見い出している。よって姓名、地名、企業名といった固有名詞について、かな見出しとID IOM記号との併用はかなりの効果を出している。優先順位とカバー率をそれぞれ横軸、縦軸にとって同じように図表化してみると、やはり同じ様にかな見出しとID IOM記号、さらに頻度情報を加えると、より少い語彙（上位の優先語彙）で90%以上のカバーが可能であるとみられる。（図3-7-9）

表 3 - 8 検索法の比較

最後にファイル構成について論述する。ファイル構成にあたってはその検索方法と考慮する必要がある。ファイル検索には種々の方法があるが、その比較をしたのが右の表3-8である。文書処理端末において通常辞書ファイルの記憶媒体として用いられるフレキシブルディスク装置では記憶容量が小さい。アクセスが遅いという制約から、インデックス、シーケンシャル検索法

検 索 法	アクセス回数	索引表の容量	記憶密度
シーケンシャル検索法	大	無	高度
2 分 割 検 索 法	中	中	高
ハ ッ シ ュ 検 索 法	小	小	高
直 接 検 索 法	小	無	中
ラディックス・トリー検索法	小	大	高
インデックス・シーケンシャル検索法	小	小	高

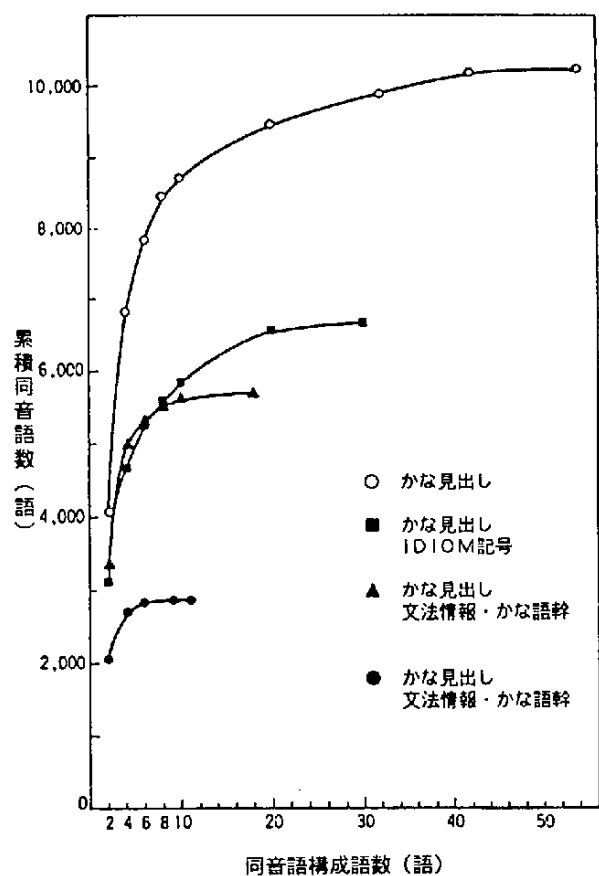


図 3 - 2 一般語の同音語分布

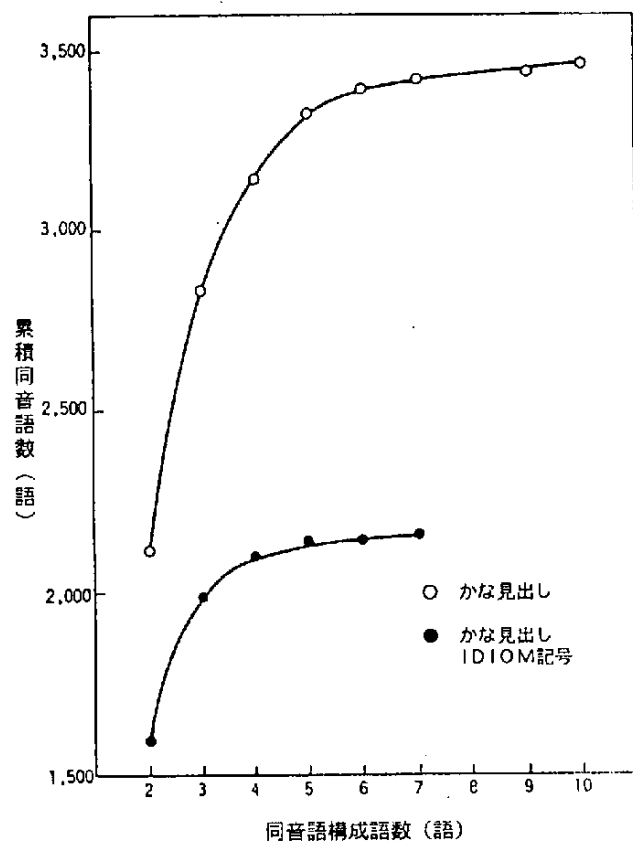


図 3 - 3 姓の同音語分布

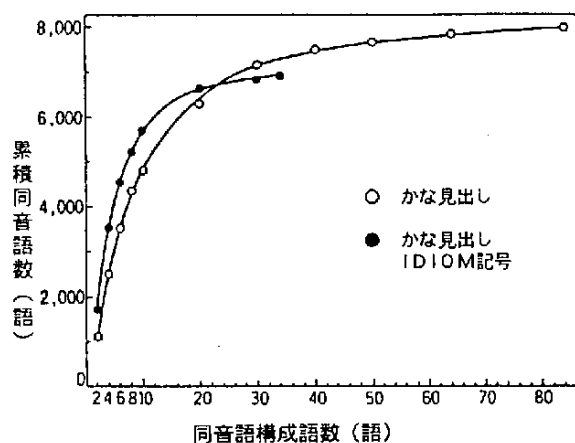


図 3 - 4 名の同音語分布

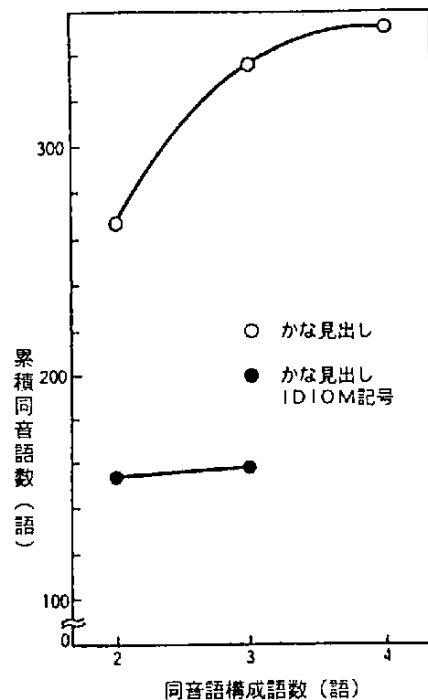


図 3 - 5 地名の同音語分布

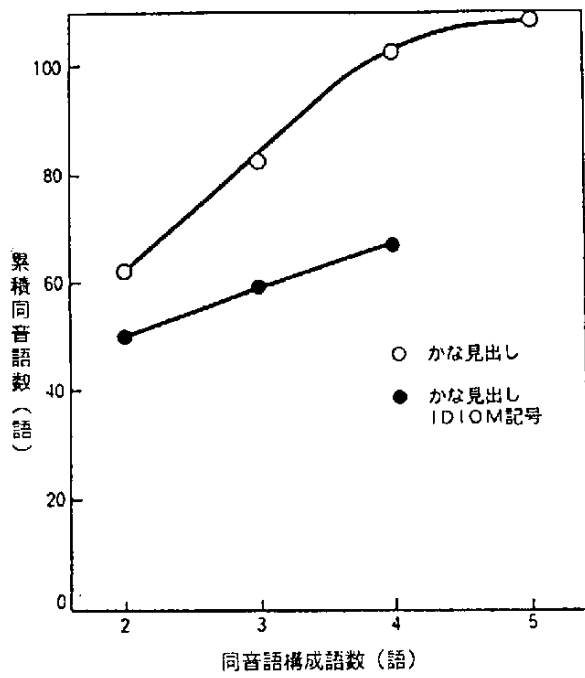


図 3 - 6 企業名の同音語分布

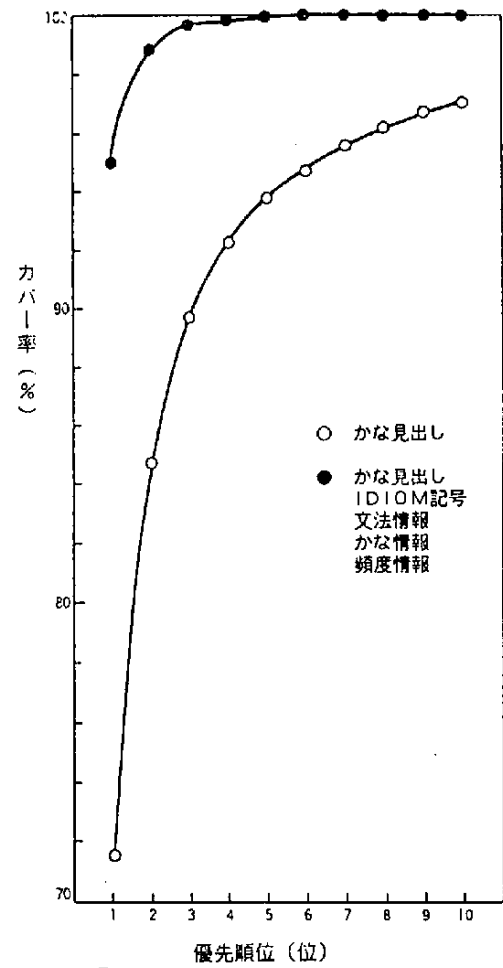


図 3 - 7
優先順位とカバー率の関係 (一般語)

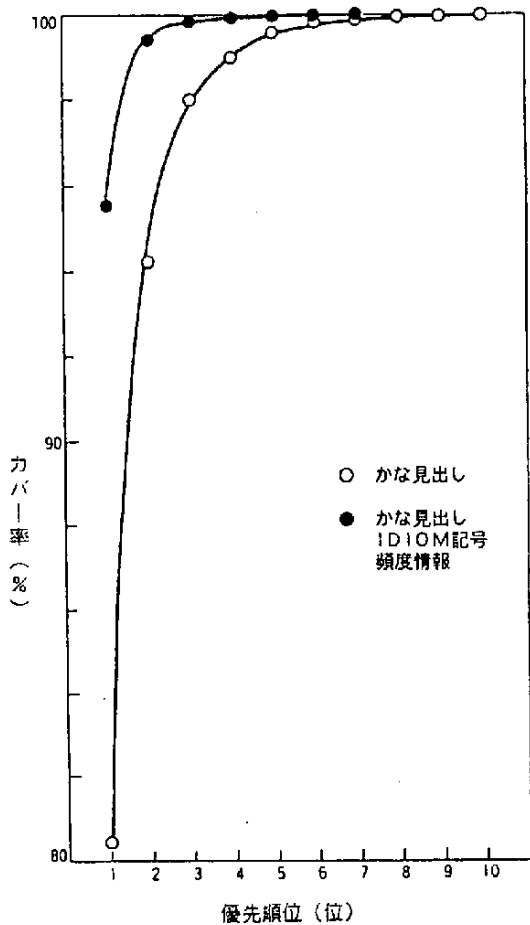


図 3 - 8
優先順位とカバー率の関係 (姓)

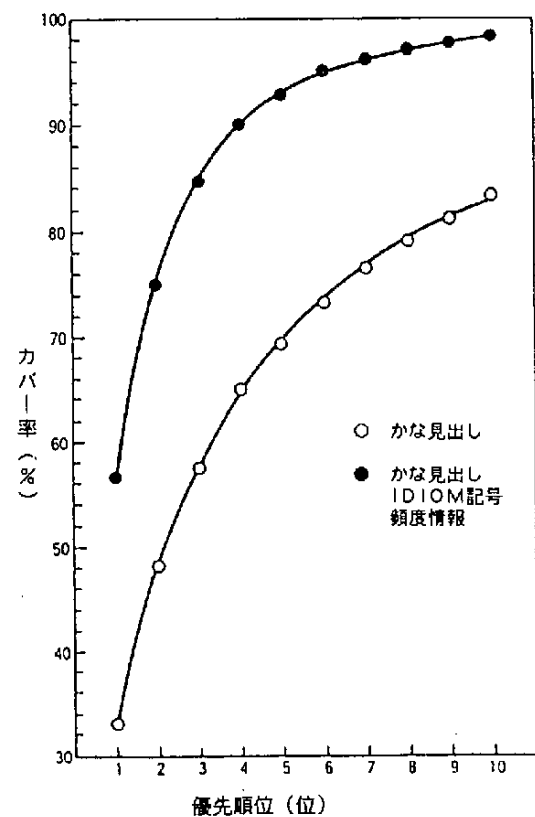


図 3 - 9
優先順位とカバー率の関係 (名)

をとることにし、ファイルを構成した。データの収録形式としては固定長と可変長があり、一長一短である。

そこで更新が頻繁に行われるであろうソースファイルには固定長、実装する場合には0.42MBまで容量を圧縮できる可変長をとることにした。

今後の課題としては、追加・修正などの更新が容易に行える辞書ファイル維持管理法の検討を挙げている。

- (5) 文献5) 菅宮和夫：J I C S T 科学技術用語シソーラス — 今後の問題，情報管理 vol. 23 No.11 pp.1011-1016(1981)

(a) オンライン時代における利用のパターン

オンライン情報検索の利用パターンは、エンドユーザーが1つの端末から多種多様なデータベースを利用できるとともに、エンドユーザーは、データベースに熟知したドキュメンテーション部門の人から情報検索に不慣れな一般の人々に広がっていくと思われる。

(b) 今後の用語コントロール

Ⅰ コントロール方式かフリーターム方式か

① 統制言語方式

長所；検索効率が低い（特に再現率が高い）。

；検索方式の作業が簡単（但し、シソーラスをよく知っていることが必要）。

欠点；科学技術の進歩に伴いシソーラスのメンテとバックファイルの更新が必要となる。

；会話モードのオンライン検索に対して効率的でない。

；ディスクリプタへの変換が必要であり、高度な専門分野への知識が必要とされる。

② フリーターム方式

長所；索引作業が容易。

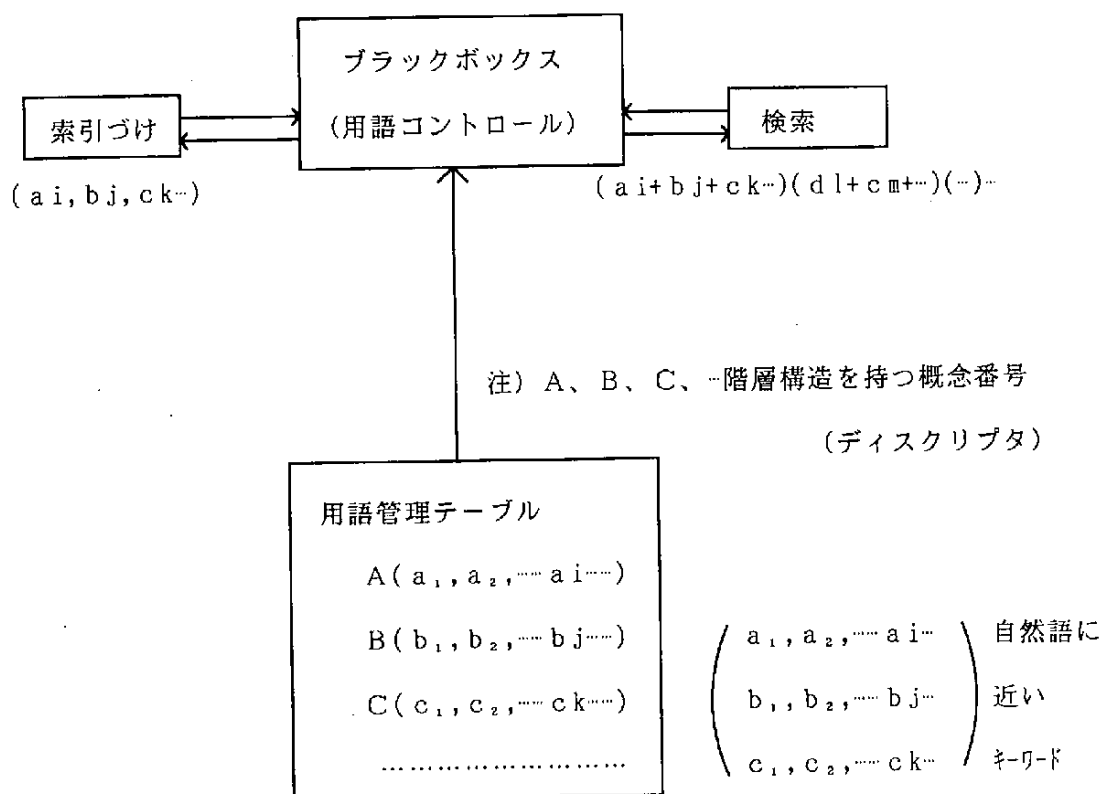
；特定の主題でも検索が可能。

欠点；同義語と想定されるフリータームをすべてORで結ぶなど検索式の作成が困難。従って検索効率（特に再現率）は、あまり期待できない。

② ブラックボックス方式

入り口（索引づけ）と出口（検索）は、自然語に近いキーワードで操作し、電算機システムでそれらのキーワードのコントロールを行うシステムである。

図 3-10



この手法は、思考実験のレベルであってこのシステムを開発するには、言語処理に関する高度な開発とともに、自然語の網羅的収集及びその個々に対応するディスクリプタとの関係づけなど膨大な作業量が発生すると思われる。

③ 多言語シソーラス

今後のオンライン検索は、任意のデータベースに対してどんな言語でもアクセスできることが望ましい。

多言語シソーラスの開発には、多国間の相互協力が不可欠であり多くの問題点もあり、第一ステップとして日英2言語のシソーラスの開発が目標となる。当面は、JICSTシソーラスの用語を英訳して電算機の利用管理テーブルに入れ活用する方法をとる。

この場合の具体的な利用方法は、入力段階では日本語ディスクリプタ→英語ディスクリプタの一方通行的自動変換にとどめ、蓄積の段階では、英語の用語については、アップホスティング機能はロックしてしまう。そのため制約はあるが、（上位語による包括的な検索ができない）英語による検索ができるようになり、そのメリットは大きい。

(c)当面の課題

① 準ディスクリプタの採用

シソーラスのディスクリプタのみでは、科学技術の進歩に対応した用語に追いつかず、いわゆる準ディスクリプタの採用が必要となった。

② 準ディスクリプタの整理と活用

準ディスクリプタは、1文献に1～2個を付与すると年間40～80万個になり、これらのディスクリプタをどのように整理活用するかの方策が必要となっている。

①類積したデータを50音順、頻度順、後方一致順などで出力しシソーラスメンテナンスに活用する。

②上記リストをオンラインユーザーに提供し、オンライン検索の補助ツールとして活用する。

③ シソーラスおよびバックファイルの更新

科学技術の発展は、日常的に新技術の用語を次ぎ次ぎと発生させ、一方のシソーラスは、ある一定の時期にくぎりをつけて用語を整理し体系づけるという作成方法をとる。このためのギャップは、少しでも縮める努力が必要である。

シソーラスの更新に関連して次に問題となるのは、新旧のシソーラス間のバックファイルの更新である。新旧ファイルを対象に検索するときにシソーラスの使い分けをしないで検索できる方法を検討する必要がある。

(6) 文献6) 鳥海剛：J I C S T 科学技術シソーラスタームチャート 構成・機能・メンテナンス, 情報管理 vol.25 No.9, pp.755-766 (1982)

(a) タームチャートとは

シソーラスの中で各ページに分散している互いに関係の深い用語を一箇所にまとめ用語

間の関係を示したものをタームチャートという。

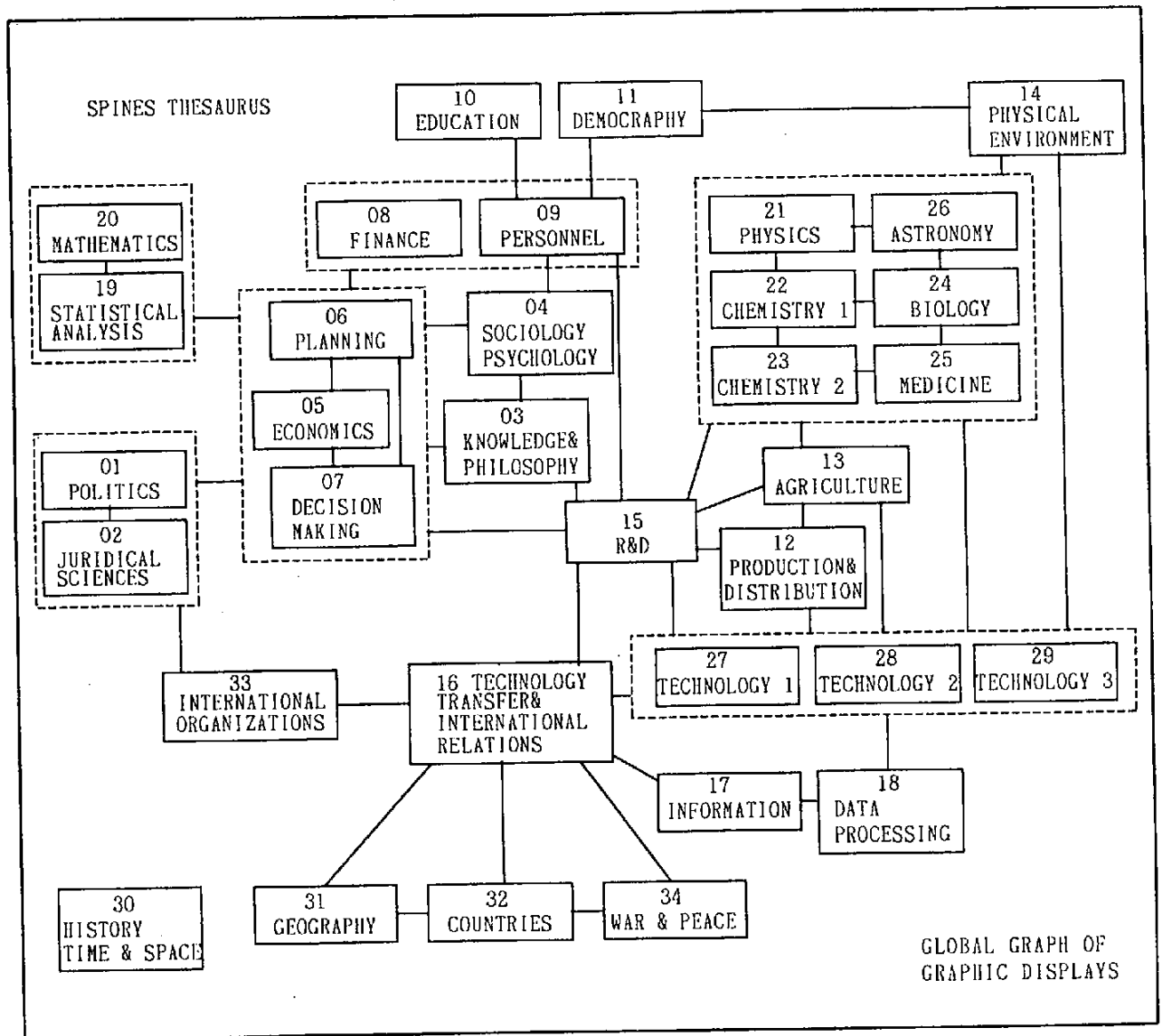


図 3-11 SPINESシソーラス タームチャート

(b) タームチャートの構成

<ディスクリプタと非ディスクリプタ>

タームチャートはディスクリプタのみで構成し、非ディスクリプタは、タームチャートのあとに主題カテゴリー別に表示とした。

<主題カテゴリー>

JICSTシソーラスは、シソーラス本体と主題カテゴリー別索引より成る。

タームチャートは、カテゴリー単位で作成された。

208カテゴリーのうち意味のない2カテゴリーを除外した。

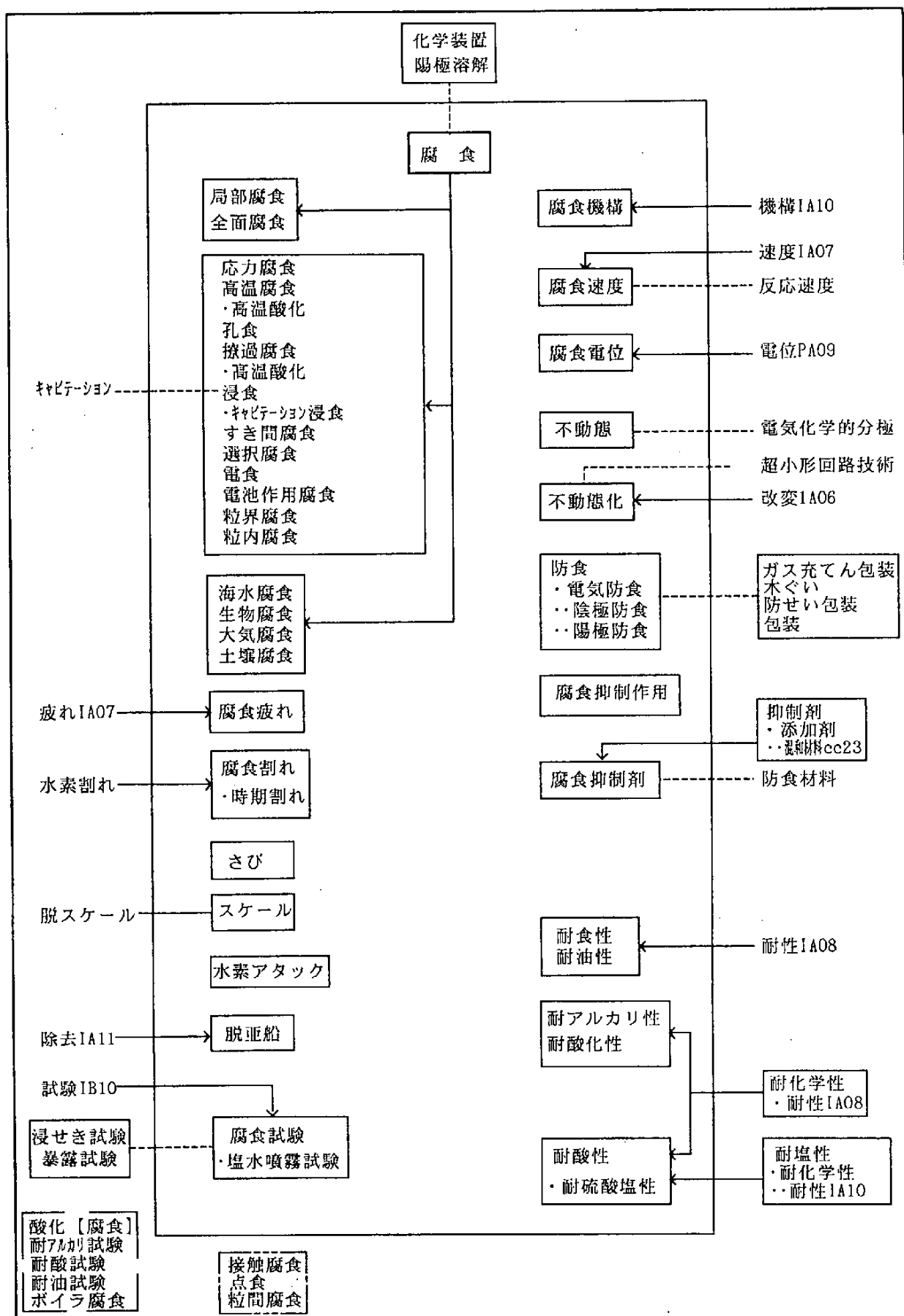


図3-12 JICST シソーラス タームチャート

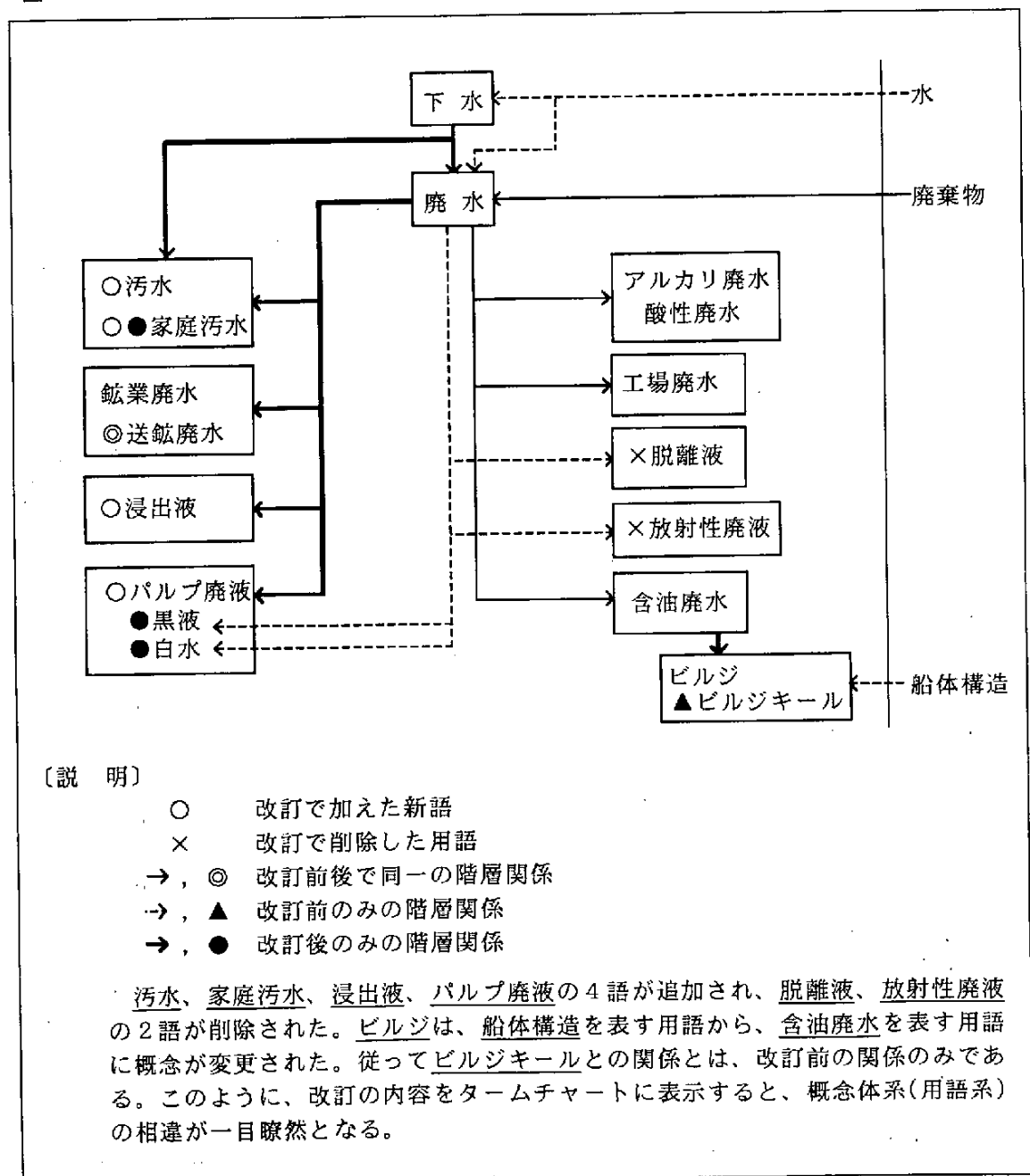
(c) 用語の配置と諸関係の表示方法

関係の深い用語の配置、階層関係にある用語の表示方法、孤立語の表示方法に具体的な工夫を実施した。

(d) ヒストリー

シソーラスの改訂に際しては、如何なる改訂がなされたか（ヒストリー）が改訂後のシソーラスに示されるのが望ましい。新ディスクリプタ、削除したディスクリプタ、降格したディスクリプタの表示が必要となった。

図 3-13



(7) 文献7) 斎藤和男 他：J I C S T 科学技術用語シソーラス 1987年版の作成，情報管理
vol.29 No.11, pp.952-964 (1987)

(a) J O I S 検索者

J O I S を検索する人の調査では、イ) J O I S 担当者、ロ) 調査担当者、ハ) 研究者・技術者、ニ) その他に層別すると、J O I S 担当者以外の利用者は、65%にものぼり、データベース利用のキーポイントは、誰でも使えるより使い易いシソーラスが必要と考えられる。

(b) シソーラスの活用

統制語と非統制語のいずれを利用するかについての調査では、イ) 主にシソーラスを使う、ロ) 主にフリータームを使う、ハ) 両方を併用するの層別の中でイ) が45%、ロ) は8%、ハ) が47%となり、ロ) は非常に少い結果になった。その理由として、「検索が確実でない」ことが読みとれる。又、フリータームを使う理由としては、シソーラスのみでは新語専門語が不十分で不足していることが上げられている。

(c) シソーラスに対するニーズ

J O I S シソーラスに対する要望の主なものとしては、イ) ディスクリプタの充実(53%)、ロ) 非ディスクリプタの充実(27%)、ハ) 新旧シソーラスの関連づけ(27%)、ニ) 関連語の充実(25%)が上げられている。

(d) 非ディスクリプタへのニーズ

非ディスクリプタの充実の必要性が高いと考えられるが、理由としては、オンライン検索の普及とともに、情報検索の専門家でない人たちの利用が急増し、シソーラスに収録されていないキーワードへもそれに伴い増えたことがあげられる。

又、ディスクリプタへの適応作業の省力化の点からも、非ディスクリプタの充実が必要となってきた。

(e) シソーラス改訂

シソーラスの改訂に伴い、新旧のシソーラスの連動による一元的な検索を可能にするため自動再索引(バックファイルメンテナンス)を実施することになった。

(f) 新語候補語の収集

新語の収集は、J I C S T データベースに蓄積されている自然語キーワードとして、準ディスクリプタとタイトル語を利用、出現頻度を用いるとともに、同一記事に索引された

分類コード、ディスクリプタを参考データとして活用するために分類コードとディスクリプタを共出現頻度順に出力した。(図3-14)

並列用統計システム		新語入力原票 (準ディス)		(85.04.10) PAGE.003614	
☐ A10	010183	*	J3	IA07	0365091
☐ B00	くん				0008
☐ B10	クン				
☐ C00	1:ディスクリプタ 2:非ディスクリプタ 3:非ディスクリプタ(USE, ANDの場合)				
☐ C10	IA07				
☐ D00					
☐ K01	BIT X				
☐ K02					
☐ K03					
CC	006 RA02	003 RB05	002 RB08	001 RA06	001 RB01
	001 XE04				
KW	006 火災			IA07	
	003 煙			GC19	
	003 家具			AB04	
	003 防火加工			IA11	

図 3 - 1 4 新語候補語入力原票

(g) 改訂作業

ディスクリプタの見直しを行った。シソーラスでのディスクリプタの数を適切な規模に保つために索引使用頻度の少なくなったディスクリプタを非ディスクリプタとするか、削除を行った。又、同一対訳語の見直し — ①外来語の音訳と意識の違い、分野による日本語訳の違いなどで同義語として扱った方が、良い場合は一方を非ディスクリプタとした。②同一英語に対し、日本語ディスクリプタが違った意味で用いられている場合はRT関係とした。③日本語のディスクリプタに対し音訳された他方が特定の概念に含まれる場合は上位一下位関係とした。

(例1) pressurization

* 与圧 USE 加圧

hierarchy

*ハイアラーキ USE 階層構造

transportation

*運輸 USE 輸送

(例2) stripper

*剥皮機 BT ストリッパ

sealing

*封かん BT 密閉

lining

*覆工 BT ライニング

(例3) acceptance

*アクセプタンス RT 受容

thermogsapy

*サーモグラフィー RT 感熱複写

surface structure

*表面構造 RT 表層構造

(h) 用語管理システムによるチェック

磁気テープに入力された改訂データは用語管理システムにより次のチェックが行われた。

- ①フォーマット・チェック
- ②使用文字種チェック
- ③見出し語の漢字-カナ変換によるフリガナ・チェック
- ④用途区分と関係子の矛盾チェック
- ⑤見出し語同形チェック
- ⑥フリガナ同音チェック
- ⑦関係語文字列チェック
- ⑧関係矛盾チェック

(i) シソーラス原案の点検と修正

用語管理システムでは、機械的なチェックであり完全ではないので、各種リストを出力し、確認作業を行った。特に新語追加は各カテゴリ単位に用語関係付を行ったため、見出し語の語尾文字列からの漢字コード順にソートしたリスト(図3-15)を用いて、カテゴリ間の用語関係付けの漏れを調査した。

(j) バックファイル・メンテナンス

旧索引データコンピュータにより自動的に新シソーラスへ再作成することにした。再

拡張用語統計システム

見出し語リスト（後方ソート順）

超音波照射	P A 12
	BT 照射
電磁波照射	P A 12
	BT 照射
食品照射	L S 71
	BT 食品加工
部分照射	P A 12
	BT 照射
外部照射	P A 12
	BT 照射
内部照射	P A 12
	BT 照射
低線量照射	P A 12
	BT 照射
入射	I A 07
プラズマ入射	P A 06
ビーム入射	P A 12
	BT ビーム技術
すれすれ入射	P A 05
	BT 電磁波入射
中性粒子入射	P A 06, P A 12
	BT ビーム入射
	BT プラズマ入射
電磁波入射	P A 12
日射	G C 19
直達入射	G C 19
	BT 日射

図 3 - 1 5 見出し語後方ソート順リスト

索引の方法は①見出し語・フリガナの変更、②階層関係の変更、③旧シソーラスでの削除語の処理、④非ディスクリプタへ変更した語の処理、⑤準ディスクリプタで索引されてい

た新語の処理の順に行い、旧シソーラスで索引されたファイルも新シソーラスを用いて検索することができることになった。

- (8) 文献8) H.ミリ (Hafdh Mili),ロイ.ラーダ(Roy Rada): マージングシソーラス 原理と評価: Merging thesauri: Principles and Evaluation, I.E.E.E. transaction on Pattern Analysis and Machine Intelligence vol.10 No.2, pp.204-220(1988)

階層構造を持ったシソーラスは分類用語とは上位概念用語、下位概念語あるいは同義語の様なその他の関連語からなっている。ほとんどの情報検索システムはドキュメントに対し索引用語集としてシソーラスを使用する。

情報検索システムに含まれる知識はおびただしく多い。それはシソーラスにある何万もの用語やデータベースにある何百万件ものドキュメントにより反映される。また、情報検索システムを維持することは通常は労働集約的な仕事であり、常に新しいドキュメントが索引づけられ、データベースに蓄積される。それゆえに、情報検索の研究者はシソーラスの構築とメンテナンスのプロセスを、また、ドキュメントの索引付けを自動化する方法を開発しようと努力しつづけて来た。

近年の関心事はドキュメントの分類と検索を改善するためにAI技術を使用する方向に向かっている。それは、マン・マシンインターフェース、知識の表現・処理、および知識の獲得に関するものである。

情報検索システムに含まれる知識をいかに表現し、また推定するかという問題は困難であり、自然言語処理技術が研究されてきている。

(a) 我々のアプローチ

我々は理論に基づく語義ネットを使いながら索引づけと検索の方策を研究して来た。我々が使った語義ネットはノードが分類語がセットで非常に良く記述されていて、リンクは“上位概念語”の関係である。

この論文で使用した『DISTANCE』アルゴリズムは理論をコンピュータ実行したものであり、我々は二つのセットの用語間の位相数学的距離と概念的距離が一致すると仮定する。

以前の実験 (1986, Machine Learning: Expert Systems and Information Retrieval) では二つの距離が近い相関関係を示している。

我々のインデキシング・アルゴリズム『INDEXER』はドキュメントを索引づけするためのシンプルな方策を使っている。基本的な方策は、ドキュメントのタイトルを構成している言葉をシソーラスから選び分類語を当てはめることから成っている。

この基本方策は増加の結果として新語がシソーラスに追加される時にわずかに修正される。この論文では同一の分野の知識を持った他のシソーラスをマージングすることによって、シソーラス（語義ネットをモデルとした）を増大させることに焦点をあてている。

この論文で提案したアルゴリズムはDISTANCEに影響を与える方法でマージされたシソーラス間での構造上の違いを利用した。

我々の仮説はマージされたシソーラスはオリジナルのシソーラスより良い構造を持つだろうということであった。

(b) 矛盾解決をしたマージング

MeSH(Medical Subjects Headings), SNOMED(the Systematized Nomenclature of Medicine), C R C S(the Computing Reviews Classification Scheme)のマージングの実験を行った結果、我々は我々のアプローチに矛盾解決の必要性があることがわかった。

2つのシソーラスをマージングするキーステップは、2つのシソーラスの間で類似の用語を見つけることから成っている。

前の実験でのアルゴリズムにおいては、2つの用語は同じ言葉の語幹を持っていれば、類似としている。このことは限定されないばかりか、類義語と同音異義語の区別ができない。それ故、我々はそれらの類似性を確めるために用語についてのより一層の知識の活用を必要とする。我々は関係語をBT's(上位概念語)、NT's(下位概念語)、RT's(関連語)、ST's(類似語)の4つに限定する。

我々のアプローチの目的のためにインプットされた2つのシソーラスはツリーとして扱い、ターゲットとしてのサブツリー-Targetとリファレンスとしてのサブツリー-Referenceとに区別されている。マージングは類義語ルートのノードを持ちながらサブツリー上で行われる。マージングは3つのステップ(図3-16)で行われる。即ちStep 1(インプット)、Step 2(シソーラス間の類似性の同定)、Step 3(Step 2により発生する矛盾の解決)である。

設定されている一方、別のシソーラスでは“パターン認識”が“A I”のNTと設定されているように、“A I”が“パターン認識”をBTとNTの両方の関係を持つことになる。

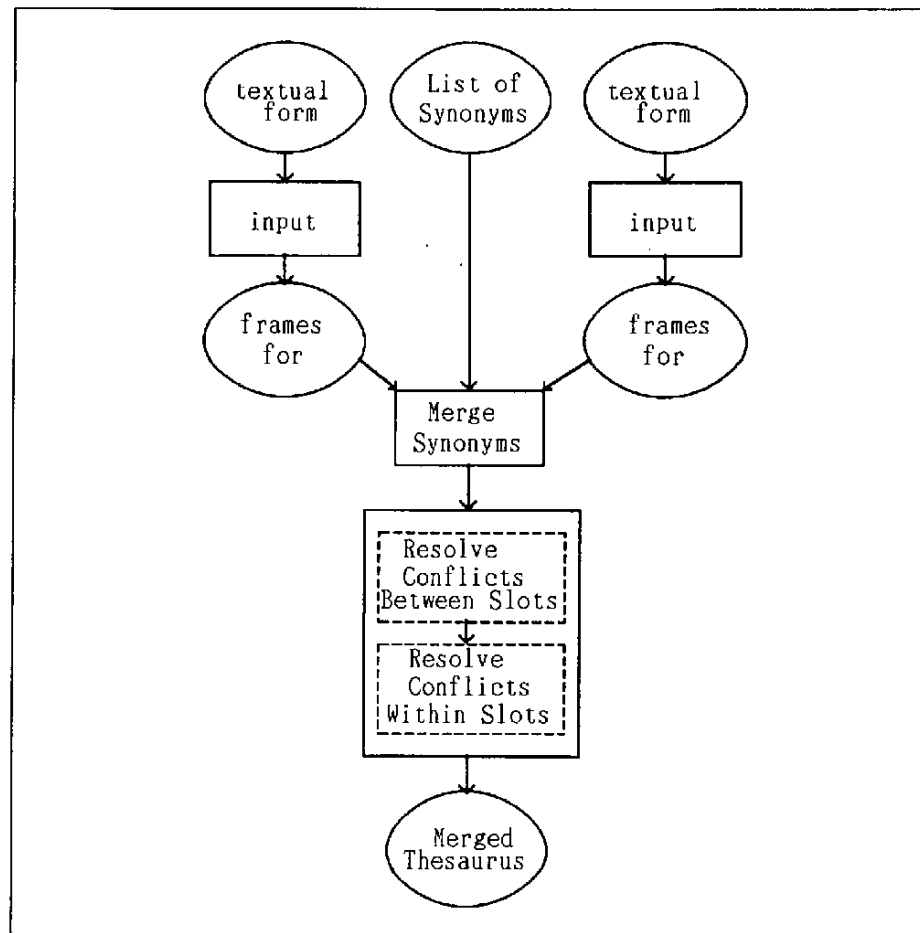


図3-16 flowchart for merging conflict resolution

この解決のルールは① RT-BTの矛盾：RTとBTの両方の関係にある時は、BTのリンクを削除する。(図3-17(a))、② RT-NTの矛盾：NTのリンクを削除する。(図3-17(b)) ③BT-NTの矛盾：BT,NTの両方のリンクを削除し、相手の別の親とBTリンクを結ぶ。(図3-17(c))

第二のタイプの矛盾は、例えば用語taが2つのBT's (親)を持っているという矛盾である。その解決策は、階層上より高いBTのレベルものの関係を削除することである。

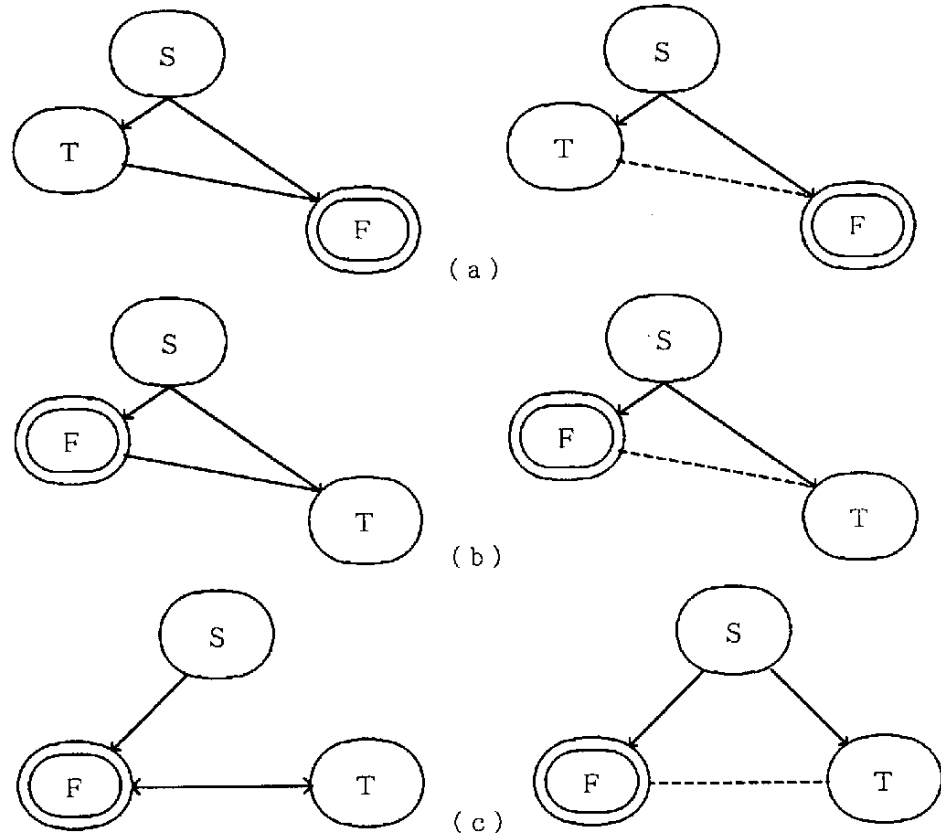
(c)結論

知識ベースがインテリジェントシステムのキーコンポーネントであり、これらの知識ベースの構築やメンテナンスの難しさは、情報化時代発展を妨げる主要因となっている。

我々の研究グループはドキュメントのためのインデキシング(INDXER)と検索(DISTANCE)これらの矛盾は2つの異った方法により解決する。

第一のタイプの矛盾は、例えば1つのシソーラスで“パターン認識”が“A I”のBTとの知識ベースとしてシソーラスを使い2つのコンピュータアルゴリズムを考案した。

このコンピュータアルゴリズム (INDEXER,DISTANCE) により類似の課題で、人手による



Conflict resolution between slots. The arrows represent NT links. The left-hand and right-hand sides are, respectively, before and after conflict resolution. The dashed lines are the removed links. (a) Conflict resolution in the case of an RT-BT conflict. F has T as parent and sibling. RESOLVE-CONFLICTS deletes the parent link. (b) Conflict resolution in the case of an RT-NT conflict. F has T as child and sibling. RESOLVE-CONFLICTS deletes the child link. (c) Conflict resolution in the case of BT-NT conflict. F has T as parent and child. RESOLVE-CONFLICTS deletes both links and makes T a child of F's other parents, i.e., makes T a sibling of F.

ものと比較した。また、増加しないシソーラスを越えて、意義ある改善効果をもたらした。

これは、我々の新しいマーキングアルゴリズムによるものである。

より役に立つには階層構造を持ったシソーラスの矛盾を知的に解決するには、概念と関係についての人間の知識を利用する必要がある。

一つの用語が同じ親と子を持っているために兄弟関係が創られる時、それぞれのシソーラスでのこの兄弟と関係する他の用語に何が起こるだろうか？この種の仕事は知識ベースをマーキングする理論を与えるべきである。

- (9) 文献9) 岸田和明 他:シソーラスの比較評価 —「概念体系の提示」の性能を中心に, 情報の科学と技術 vol.38 No.10, pp.565-572(1988)

シソーラスは情報検索のツールの1つであるが、これまで評価事例はあまりない。ここでは検索システムから切り離した形で複数のシソーラスの比較評価を試みている。

評価する前に、その基準としてシソーラスの役割について検討し、大きな役割として次の2点を掲げる。

①統制語彙への翻訳ツールとしての役割—これは「シソーラス」という語が情報システムに登場した頃から唱えられていた。文献中の概念を統制語彙にいかに正確に・容易に変換するかが重要となる。

②概念の捉え方を提示する役割—シソーラスが特定の情報システムの中で、重要概念の抽出に一つの視点を与えるような概念体系、或いは知識体系を提示するものとする。

こうした役割は、検索者が主題とする分野にあまり精通していない場合、検索対象とする事項、物に対して明確なイメージを捉えていない場合、又は検索者と索引語作成（決定）者との同概念に対する捉え方にずれがある場合に有効である。

以上2点のうち、後者である「概念の捉え方を提示する役割」を重視すると、シソーラスが提示する概念体系が検索者にとって理解しやすいものであるかどうか、その提示の性能が重要になってくる。ここではその点に注目し、評価の視点とする。

又、評価の立場は検索者側に立つて行う。

対象となるシソーラスは現在索引作業に使用され、維持管理の実績があるものとして、ERIC, INIS, INSPEC, JICST, MeSH, 日経, ROOTを選んだ。

評価ポイントは（表3-9）の項目にあるが、特にスコープノート、階層構造、シソーラスの作成方針などの解説が重要であるとする。スコープノートは概念は文章で表現されることから、階層構造は語彙の上位・下位関係でそのシステムの概念の捉え方を知り得ることからその重要性が生じる。評価結果は前出の表3-9にまとめている。

シソーラス構成要素については、ROOTにファセット・インディケータの表示機能があり、他より優れていることをあげ、ERICに語順リスト・階層リスト間の参照ができないこと、INISでは階層リストがないために構成の一覧ができないことの問題がある。

階層構造としては唯一MeSHにフローティング・タームがない点と、対象としたシソーラスがすべてラティス構造である点を評価する。

エントリーの構造を見るということは、語順リストの各語彙の表示を見ることで、スコープノートが重要となる点である。ERIC以外はスコープノートが設定されてはいても

その機能は、概念体系の提示という点で不十分である。語間関係の表示はBT,NT全段階表示が望ましい。語の更新情報は、概念体系の時間的推移に重要な位置を占める。限定句は、同形異義語の処理以外に語を促える一つの視点を示すと考えれば重要である。しかし、実際に検索するには煩わしく、利用しやすさを重視するならばない方がよい。

維持・管理については筆者らがかなり主観的に断定したものであるが、各項目とも利用者にとって重要な情報源である。MeSHは用語の収集方針の説明、JICSTは使用法の解説の点で特に優れていた。

全体を通してみると、ERICとMeSHがほぼ全般で高い評価を得ており、概念体系の表示という範囲では優れたシソーラスといえる。ERICは語順リストと階層リストの間に参照機能がないという欠点はあるが、スコープノートは他に比べて特に充実している。MeSHはスコープノートが数少ないが、語の意味を規定する他の工夫が様々になされている。JICST、日経シソーラスも、スコープノートの不満足という点を除けば「提示」機能はかなりあるといってよい。注目すべき点として、ROOTは階層リストにファセット・インディケータを持つという長所がある。

今後の情報システムにおいて、現在研究中の自然言語処理などの「知識表現」の問題とも絡み、シソーラスの提示機能はより重要になると予想される。探究すべき問題への手がかかりとして、具体的には、どのような実体のない概念を表現するか、どのように統制語彙を規定すべきか、あるいは上位・下位関係以外の統制語間の関係をRTとしてひとまとめにしておくだけでよいか、などがある。

〈問題点〉

本論中に「今後の課題」の中で掲げているとおり、本論は現状把握と比較調査を行なっているのであって、「提示」の望ましい方法やその性能を向上させる改善について具体的言及はない。また、「概念体系」としてのシソーラスの捉え方が、充分であるかどうか、評価ポイント（表3-9の項目）を細分することはできなかったかも気になる点である。

特に表で全体の結果を示しているが、「○」「×」でその機能の有無を表している場合、その程度が問題となる。例えばスコープノートの評価について「ほとんどない」「不十分である」とどちらであっても表の上では「△」である。そうした点を明示できれば一覧してどういう使い方にはどのシソーラスが適当かということまで含めて評価できたのではないと思われる。

表3-9 シソーラス評価ポイントと評価結果

○…有 ×…無

	ERIC	INIS	INSPEC	JICST	MeSH	日経	ROOT
a. シソーラス構成要素							
階層リスト	○	×	○	○	○	○	○
語順リスト	○	○	○	○	○ 2種類	○	○
両リスト間の参照	×	—	○TT参照	△	○	○	○
その他	ディスクリプタのKWIC索引あり			図式表示(タームチャート)あり	補助資料あり	補助資料あり	
b. 階層構造							
フローティングターム	○	○	○	○	×	○	○
構造 ツリー/ラティス	ラティス	ラティス	ラティス	ラティス	ラティス	ラティス	基本的にラティス(選択できる)
c. エントリーの構造							
スコープノート	○	△	△	△	△	△	△
同形(音)異義語 あるいは限定句	() で表示	× 1)	() で表示	[] で表示	() で表示	同音識別 子(1,2)	() で表示
語間の関係の表示 方法	BT,NT 一段階	BT1,BT2 BT3… NT1,… 全段階	BT,NT 一段階 TT(最上位語)	BT,NT 全段階	階層コード のリストのみ	BT,NT 全段階	<(BT) >(NT) 一段階
分類コード	×	×	○	○(主題 分析-)	○	○	○
語の更新情報	○(語の採用 年月・削除 年表示)	→(新しい 語) 語の追加 年月の表示 あり	DI(語の 採用年月 PT(以前の 語))	×,ただし 一覧はあり	○(採用 年,追加 削除一覧 あり)	N(新しい 語)	×
索引使用頻度	○	○	×	○	○	○	×
その他						キーワード の種類 の表示が いくつか ある	**,+,*,<, *>,*-,な どの記号 もある 階層リス トに、ファ クト・イン デックス あり
d. 維持・管理							
範囲・特色の紹介	○	△	△	○	○	○	△
使用法の解説	○	△	△	○	○	○	○
用語の収集方針	△	△	△	○	△	△	△
語の更新	○	○	○	○	○	○	○
シソーラスの更新 回数	改訂11回 ほぼ1回/ 2年	1回/年	1回/2年	1975,78, 81,87年に 発行	1回/年	1985,86年 に発行	1981,85 年に発行

注1) 非ディスクリプタには、()で限定句が付与されている。

(10) 文献10) L.K.リースポーター (L.K.Rees-Potter) : ダイナミックシソーラスシステム:

『社会学経済学におけるターミノロジー的および概念的变化の計量誌学的研究ならびにダイナミックシソーラスシステムの設計への適用』; DYNAMIC THESAURAL SYSTEMS: A BIBLIOMETRIC STUDY OF TERMINOLOGICAL AND CONCEPTUAL CHANGE IN SOCIOLOGY AND ECONOMICS WITH APPLICATION TO THE DESIGN OF DYNAMIC THESAURAS SYSTEMS, Interation processing & Managemanto vol.25 No.6, pp.677-691(1989)

シソーラスは情報検索の分野でインデクサーやサーチャーが使うためのディスクリプタを用意するものである。その問題点は、シソーラスのメンテナンスとアップデートにある。

この事例は、引用、相互引用の解析と引用関係解析を利用してシソーラスのメンテナンスとアップデートをするためのそのメカニズムの研究をするものである。この研究の目的は、シソーラスのアップデートとメンテナンスの効果的な進め方を開発することにある。

この方法は、レファレンスや脚注をつけているフルテキストデータベースに適用されることができる。

その方法では、新しいタームと新しい引用関係をシソーラスに反映するために3つのモデルを開発した。

①引用の同一化モデルこれは、フルテキストデータベースのリファレンスの中から強い引用、相互引用のドキュメントを発見する。

②文脈の同一化モデルこれは、テキストのなかから実際の引用を発見しその文脈を表示する

③シソーラスのメンテナンスモデルこれは、新しいフレーズを認知し、そしてその新しいコンセプトを明記し、かつ現行のフレーズのコンセプトとの関係を明記する。

データは、社会科学のインデックス(26の経済学、社会学の専門分野)から採用した。候補のシソーラスの用語は、これらの分野の文書の引用関係から半自動的に作成された。専門家は、これらの用語が専門分野に適用できるものであると結論づけた。

このシソーラスの用語は、スタンダードな標題リストによるターミノロジーにたいし良好な比較ができた。

この研究での結論の一つは、用語の発生と用語の変化に関する方法論である。一般的に、ターミノロジーは、専門分野のコンセプトを反映した文書の文脈を使う。(つまり、強い引用と相互引用を持つドキュメントは、コンセプトシンボルや強いオリジナリティを持つ

ことができる。)

この方法論は、経済学、社会学の専門性を表わすオルターネイトターミノロジーのための豊富なソースを提供する。この事例は、全ての専門分野においても同様である。

- (11) 文献11) 岡野弘行: J I C S T 科学技術用語シソーラス, 情報と科学と技術 vol.39 No. 12, pp.558-566 (1989)

J I C S T シソーラスは、科学技術分野を対象として1968年から全く新たに作成着手された日本語総合シソーラスである。

1969~70年にキーワード索引の標準化、71年に第一次標準キーワードリストの作成、73年に暫定版、75年に第一版を完成。その後78年、81年、87年と版を重ね、92年頃を目処に改訂準備中である。

(a) シソーラスの構成・構造

総語数では73年版の31,752語から、削除と追加を繰り返し87年版では48,196語に拡大、75年版では用語選定規則の改訂、81年版では階層関係の思想変更が行われている。ディスクリプタは73年版の28,196語から87年版の38,407語に拡大、抑制が必要とされることから現状が限界と思われる。非ディスクリプタは多いほどよく、73年の3,556語から87年版の9,789語に拡大している。

用語については次の通りである。

①見出し語には、非ディスクリプタである場合のみ「*」記号にて示し、

②後に () 内にてフリガナを振る。

③見出し語の次行に、14分野207カテゴリ分類された主題カテゴリコードを設定する。但し「目安」程度であり、現在は改訂作業のコード的役割を担っている。

④カテゴリコードの後に索引頻度を示す。

⑤同義語、準同義語、類義語の表示は、見出し語が非ディスクリプタの場合は「USE」・「AND」、ディスクリプタの場合は「μF」点「μF+」を優先的關係として頭に設定している。さらに、非ディスクリプタ見出し語の検索同義語、同形同義異音語の優先關係を補追するため、見出し語の方の頭に「*」に代え、「☆」記号にて表示する。(87年版より)

⑥階層關係にある用語は、「包摂關係」である時は「NT」、さらにその用語の下位概

念には順次「・」表示で列記する。「部分－全体関係」であるときは「B T」を頭に設定する。「部分－全体関係」は、

1) 地理的位置関係 2) 生体の組織・器官 3) 学問分野。

イ) 広義の一般用語 (81年版より) ロ) 多くの分野で共用される専門用語 ハ) 特定分野に所属させるべきでない／所属させることが不適当な専門用語に限定し、87年版からは共通語尾を有するものも追加した。

⑦同義・階層以外の密接な概念関係には、ディスクリプタ間の関係に限り、「R T」にて表示する。

⑧同形異義語の区別には、用語の後に見出し語限定句を付加し、

⑨同音異義語の区別には、振りがなの後にフリガナ限定句を付加する。

⑩分野や時代によって概念範囲が変わらない定義は、スコープノートとして必要に応じてディスクリプタに付与した。

化合物・元素については本体は上記①～④迄とし、他は別冊による2分冊形式としている。また、本体の補助資料として、体系・順列・主題カテゴリ別・英訳主題カテゴリ別・英訳順列の各索引・主題カテゴリ別・英訳各シソーラス・日英対訳・日英対訳各リスト・タームチャートがある。但し、体系・主題カテゴリ別各索引は、改訂作業用であり、順列索引は第3版迄、タームチャートも一部しか作成されていない。

(b) 準ディスクリプタ (自然語)

人手による自然語索引として「準ディスクリプタ」索引を実施している。但し、「準ディスクリプタの付与は、それに対応する上位概念または類似概念を表すディスクリプタを付与することを前提」とする規定付である。しかし、現在平均1語／文献が付与されている。

	延 べ 語 数	異 な り 語 数
86/8-87/7	654, 739	99, 620
86/8-89/7	1, 284, 350	173, 797

シソーラスでは専門用語や新概念の用語が不足するため、フリータームの使用度が8%、シソーラスとの併用使用度が45%にのぼる。

準ディスクリプタはシソーラス改訂データとしても利用されており、87年版作成時には頻度5以上の9,262語中4,927、標題抽出語は頻度11以上の18,774語中2,111を採用した。

(c) 問題点と今後の課題

新旧シソーラス間での対応・関連付け（ヒストリー）及び、バックファイルメンテナンスの問題については、準ディスクリプタも含めて既索引語に対する文字列より、登録番号による強制置換の手法しかない。

シソーラスに未登録の概念用語、あるいは類似の概念が索引者／検索者によって、対応するディスクリプタにバラつきを生む。この問題解決のため、J I C S Tでは、自動変換・参照辞書の搭載、あるいはA I 技術が当然となった今日、自然語－統制語変換辞書の形で採用する予定である。変換辞書としての索引支援システムは開発中であり、検索支援はJ O I S -Ⅲとして90年4月に搭載する計画である。索引の標準化・効率化および検索の適合性・再現性の向上を旨としている。

- (12) 文献12) 村田雅胖：新聞記事情報とN K -M E D I Aシソーラス，情報の科学と技術
vol.39 No.12, pp.582-587 (1989)

新聞記事のデータベース化は、コンピュータに依るところが大きい、すべて自動化するまでには至っていない。たとえば、キーワードの付与は、インデクサーが手作業で行っている。主題分析を行い、最も適切と思われる用語をシソーラスから選び出し付与している。また、記事の内容がシソーラス用語だけでは十分表現できない場合、フリータームで補うことにしている。キーワードの付与をコンピュータにより自動化しようという研究がなされ、一部ではすでに実用化しているところもあるが、当社でもキーワード自動切り出しの研究を進めているところである。

(a) シソーラス作成の経緯

当データベース構築構想は、新聞製作のC T S 化研究に伴い、除々に本格化し、1980年頃からシステム開発を進め、同時にシソーラス設計にも取りかかった。

当初は、既成のシソーラスを用いることも検討したが、新聞記事の特殊性、収録対象分野の違いから、独自のシソーラス作成を目指すことにした。まず、自社の新聞および出版物を使い、製品関係、技術関係、産業・経済関係および性質や現象を表わす用語、固有名などを収集した。

次に、収集した用語に、一定期間の頻度チェックを行い、索引語（ディスクリプタ）として適当かどうかの検討を加えた。

採用した用語は、表記の統一、同義語の整理、同音異義語の識別、階層関係の設定、カテゴリへの分類などの作業を行い、コンピュータへ登録した。その結果、総語数約9,600語の初版シソーラスが完成した。

(b)シソーラスの改訂

初版作成以降、フリータームとして別途蓄積してきた用語を中心に、索引語として使用できるか、同義語が既に収録されていないか、さらに、既登録索引語の頻度のチェック、カテゴリの見直しなどの作業を継続して行い、約3年を経過して、改訂（第2版）を行った。改訂の要点は次のとおりである。

①見出しの語の追加

経済、金融、バイオ、情報などの分野を中心に約2,100語の見出し語を追加した。このうち、索引語は約1,900語で、フリータームとして使用頻度の高い用語が中心になっている。非索引（非ディスクリプタ）は約200語。フリータームの中で、同義語を整理、一方を非索引語として索引語へ誘導する措置をとった。

②新カテゴリの設定

キーワードの収集、カテゴリへの分類を進めていく過程で、従来のカテゴリへの分類を進めていく過程で、従来のカテゴリ分類では不都合が生じてきた。第2版では、カテゴリの大幅な改訂は控えた。

③索引語の表記変更

索引語の表記の見直しを行い、一般的によく使われる用語へ変更した。

④索引語の削除

索引語のうち、使用頻度が極めて低い用語、概念がはっきりしない用語、同義語の整理の結果、不要と思われる用語を削除した。

(c)シソーラスの概要

現在用いているシソーラスは、索引語約11,200語、非索引語約530語を収録している。

索引語、非索引語は、それを統合して50音別索引、カテゴリ別索引に分類している。50音別索引では、見出語（索引語、非索引語）が数字、アルファベット、日本文字の順に配列されており、用語辞書的な役割をはたしている。また、カテゴリ別索引では見出し語を150語のカテゴリに分類、用語間の階層関係を表示すると共に、非索引語も付与している。同一階層内での配列は、50音別索引と同じである。

同義語を整理する方法として、索引語としない語（非索引語）に「USE～」を付与して、同義語の中でどれが優先するかを指示している。また電子計算機－コンピュータなどのように、一般的によく知られている語、およびよび合併事業－生産－合併のように前方が一致する語などは省略している。

カテゴリ別索引では、検索者が当該キーワードを迅速に見つけ出すことができるように、一部の索引語を複数のカテゴリへ重複して掲出している。「センサー」を例にとれば、電子部品（DJ）、計測機器（FC）の二つのカテゴリに指出している。（図3-18）

シソーラスのコード体系は索引語、非索引語のすべてに8けたのコード番号を付与し、アルファベット2けたはカテゴリを表しており、数字6けたはその用語の所在地を示している。（図3-19,20）

動いている情報を正確にとらえ、把握するためには、その検索システムの精度の向上が望まれる。そのためには、特にシソーラスのメンテナンスが必要不可欠の作業となり、第2版で初めてそのメンテナンスを行ったが、今またその見直しの必要にせまられている。

図3-18 他カテゴリへのリンク例

センサー（センサー）	
DJ	FC
UF	検出器
NT	圧力センサー
	・圧力分布センサー
	・微小差圧センサー
	・イオンセンサー
	・位置決めセンサー
	・イメージセンサー

図3-19 シソーラスのコード体系

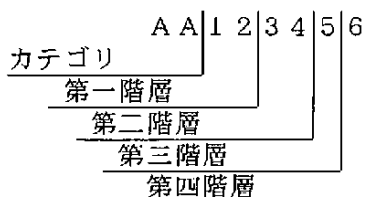


図3-20 コード番号例

a.GA100000	・	住宅設備機器
b.GA100100	・	温水器
c.GA100200	・	換気扇
d.GA100300	・	給湯器
e.GA100310	・	ガス瞬間給湯器
f.GA100320	・	石油給湯器
g.GA100400	・	サッシ
h.GA100410	・	アルミサッシ
i.GA100420	・	スチールサッシ
j.GA100430	・	プラスチックサッシ

(13) 文献13) 荒井重信 中村正則：技術データベースの開発，情報科学技術研究集会発表論文集 vol.24, pp.87-94 (1988)

一般的に、情報をデータベース化し管理運用していくためには多くの人手と開発期間を要する。企業内データベース成功の鍵は、システム開発の効率化と管理運用の低コスト化にあり、そのためには、データベースの開発から運用まで一貫した支援システムが必要である。本論文では、抄録を含んだデータベースには不可欠なキーワード処理の課題とデータベースの種類が増加した時の課題を解決する一方策を述べる。

(a) キーワード索引の種類と課題

代表的なキーワード索引の方法として①分類コード、②統制語、③フリーキーワード、④機械抽出の方法がある。①～③の方法はそれぞれ分類表、シソーラス、用語集などのマニュアル類が必要で、キーワード付与にはある程度の熟練が必要であり、または付与者の考え方によって同一データでも異なるキーワードが付与される場合がある。④の方法は特別なマニュアルは不用で、キーワードも機械的にかつ同一の思想で付与されるため、①～③と比較して作業効率の向上およびコスト減が計られる。しかしこの方法は原文表記の違い（同義語）を意識した検索式をたてる必要があり、その点が問題である。

(b) 同義語の種類と整理方法

同義語は10種類に分類され（表3-10）、②、⑤は略語辞書や外来語辞書を利用することにより比較的容易に整理できるが、その他のものは、適当な辞書がなく自前で整理するし

表 3 - 1 0 同義語の種類

同義語は次の10種類に分類される。

- ① 意味的：計測器 測定器
- ② 略語：DB データベース
- ③ 表記の相違：OPEC オペック
- ④ 送りがなの相違：読み込み 読込み
- ⑤ 異字体表記：製鉄 製鐵
- ⑥ アクロニム表：IBM I. B. M.
- ⑦ 長音記号付与の有無の相違：コンピュータ コンピューター
- ⑧ 外来語表記の相違：硝子 ガラス
- ⑨ 四つ仮名・旧表記体：ベニス ヴェニス
- ⑩ 助数詞表記の相違：第11回 第XI回

表 3 - 1 1 同義語の整理例

【用語】 O A 【副用語】 オフィスオートメーション オフィスオートメイション	【用語】 D P C 【副用語】 運転基準出力制御運転 給電出力指令	【用語】 S I 【副用語】 シリコン ケイ素 珪素 国際単位系
【用語】 F R P 【副用語】 繊維強化プラスチック 繊維強化プラスティック	【用語】 L D 【副用語】 レーザダイオード 致死線量	【用語】 東京電力 【副用語】 東電 T E P C O

かない。本システムでは技術開発関係の用語に絞って整理を行った。（表3-11）

(c)同義語の処理方法

同義語の処理方法として次の3つの方法が考えられる。①標準語変換：データベース作成時に標準語（ディスクリプタ）に変換して格納、②同義語付加：データベース作成時に全同義語を付加して格納、③検索依存型：データベース作成時に同義語処理は一切しないで格納がある。本システムでは、現状で検索レスポンスが一番良い、②の方法を採用した。

(d)同義語処理システム

本システムはキーワード自動抽出機能、同義語辞書管理機能と同義語付加機能から構成されている。（図3-21）キーワード自動抽出機能は平和情報センターのHAPPINESSを利用した。

同義語辞書は（図3-22）に示すように、一同義語群は標準語とその同義語の組で構成されていて、同義語の数の制限は特にない。

同義語辞書管理機能として①辞書の登録・更新機能、②同義語重複チェック機能、③辞書一覧出力機能を有している。

辞書の更新は次の手順で行っている。①ある期間に付与されたキーワードのリストと累積キーワードリストの比較処理を行い、新規発生キーワードのファイルを作成する。②先

に述べたキーワードの整理方法にそってリステイングし、同義語の整理を行う。③辞書の登録・更新機能を使って辞書を会話方式で更新する。更新は、登録したい同義語群を呼び出し、大義語群が既に存在する場合はその同義語群を表示して新キーワードの追加を行う。該当同義語群が存在しない場合は、新同義語群を生成する。

表 3 - 1 2 同義語処理結果例

半導体電極を用いる太陽光エネルギー変換

半導体電極による太陽エネルギー利用法は次の3つに大別することができる。

- ①直接電力に変換する（湿式光電池）。
- ②水を分解して水素、酸素を生ずる。
- ③水の分解以外の化学変換。

このうち①は安価で効率の高い光電池を得ることができるかもしれない。②はクリーンで、応用性の広い燃料を得ることができる。③は例としてヨウ化水素の分解をあげたが、水の分解にくらべると技術的にはとっつきやすい方法であり、シリコンのようなバンドギャップの小さい半導体を表面処理によって安定化し使用できるようにすれば高い効率をもちしかも安価な変換素子がえられることになる。

半導体	応用性
半導体電極	ヨウ化水素
太陽光	シリコン
太陽光エネルギー	バンド
太陽光エネルギー変換	バンドギャップ
エネルギー変換	ギャップ
太陽	安定化
太陽エネルギー	変換素子
太陽エネルギー利用法	素子
エネルギー利用法	H
湿式	O
湿式光電池	SI
光電池	珪素
水素	ケイ素
酸素	国際単位系
化学	
化学変換	

同義語付加機能は同義語辞書を参照し、抽出されたキーワードへ同義語を付加する機能である。（表3-12）に同義語処理結果を示す。H、O、S I、珪素、けい素などは同義語として付加されたキーワードである。

(e) データ情報管理システム

データベースを利用するのに必要な情報を管理するシステムとしてT I S D M (Technology Information System D A T A / D Manager) を試作した。これは、データ辞書の機能を拡張したものである。それは、①ファイルに関する情報、②データに関する情報、③コード表と利用情報、④操作マニュアル情報、⑤利用者管理情報、⑥個別アプリケーションの情報の管理をしている。

(f) データベース利用者システム

汎用検索システムT I S (Technology Information System) の検索方法は①フリー検索方式：利用者は頭に浮かんだ技術用語で検索する方法、②テーブル検索方式：利用者が選択したデータベースの項目名を画面に表示し、利用者はそのテーブルに検索条件を入力する方法、③メニュー方式：業務処理のメニューが画面に表示され、利用者が対象業務を選択する方法の3種類を用意した。

①と②は情報検索用のシステムで利用者の経験により選択できるようになっている。

(①は初心者向き) ③は定例業務処理用の検索システムである。

(g) 今後の課題

今回は、抽出されたキーワードの横の拡がり（同義語）を扱ったが、原文に直接表れない言葉も加味すれば、検索の自由度はさらに増す。特に、文書情報の検索においては、漠然とした記憶から検索する場合が多いので、知的検索システムが必要となる。そのためには、キーワードの横の拡がりと縦の拡がりを合わせた立体的な辞書（知識ベース）が必要で、今後研究を進める予定である。

課題の2点目としては、異なるメディアの統合化である。情報の中には写真・図・表等文字以外の情報が含まれている。これらの異なるメディアを管理し、自由に組み合わせることが可能なデータベースマネジメントシステムが実現すれば、日常業務の生産性向上にも一層寄与すると考えられる。

(14)文献(14)岡野弘行 他：自然語－統制語変換辞書作成システム 情報科学技術研究集会発表論文 vol.25, pp.49-55 (1989)

計算機資源の進歩により、文献データベースの利用技術は大いに発展している。ところが、そのデータベース作成の中心である抄録作成と索引作業が専門家に依存せざるを得ず省力化が難しいため、昔日と較べて著しい差は見られない。しかし、省力化を目指し、抄録作成においては、著者抄録の徹底と標準化をはかり、それを利用するという形に移行することにし、索引作業については、論文標題や抄録などからのキーワードの自動抽出システムの開発によって、省力化、高速化という点では展望が開けつつある。ところが、後者のキーワード自動抽出は、標題および抄録中のテキスト語（自然語）が即検索語となっているため、トータル的に「自然語が有利か、統制語が有利か」という古くて新しい問題を自ら解決していない。

自然語と統制語の優劣がつけ難いならば、両者の利点を併せ持つ「自然語から統制語への交換インターフェイス」を構築することによって解決できるのではないかと考えて、その変換アルゴリズムを設計し、変換辞書の作成に着手した。

(a)開発の背景

一般に統制語索引によれば、検索結果は網羅性が保証できるといわれている。これは統制語索引がバラつかないとの神話が必要であり、熟練した索引者においても、バラつきは抑えられず、逆に、自然語索引の方が統制語のそれよりも一致する語は多いこともあるとの報告すらある。このバラつきを防ぐ手段はないかという命題のもとに、自然語から統制語への案内辞書、すなわち同義語への参照可能な辞書の作成を検討した。一方、J I C S Tでの「抄録と索引作成」という実習指導の実験から、索引の手順としては、ドキュメンテーションに不慣れな、精通していない人については、自然語で索引語を抽出させ、その後シソーラス統制語に変換する方が、バラつきがないことがわかった。さらに自然語から変換すべき統制語が案内された辞書を備えているならば、一致率は上昇し、かつ索引に要する時間も短縮可能と予想できる。

ところで、自然語から統制語への変換辞書の作成手法として共出現などの統計的手法と意味的解析手法が考えられる。統計的手法はある程度蓄積された過去のデータに対してのみ、変換辞書を作成できるわけであり、データベースを新たに作成する場合には作成できず、さらに少しでも綴りの異なる新語に対してまったく用をなさない。そのため、データ

ベースには直接関わらず、統制語辞書シソーラスに依存した変換辞書および変換システムは開発に値するものと考えられる。

そこで、シソーラスをある一つ概念体系、あるいは知識体系と捉えると、統制語すなわちシソーラス用語、およびそれらの構成する体系によって、一つの意味空間が形成され、その同義関係、上下関係が自然語から統制語へ変換する際に何らかの役割を果たすであろうと予想される。

さらに J I C S T における汎用シソーラスにおける永年の経験から、シソーラス体系と用語の語形を利用した変換システムの開発の可能性あることを確信した。

(b) 自然語－統制語変換システムの開発経緯

1240 文献 (20 件/人) について人手作業により統制語索引と自然語索引を実施し、両索引の照合 (対応づけ作業) を行った。

おのこの統制語索引の対応づけとしては、自然語に対して統制語が上位語、同義語、その他 (固有名詞のため、対応なし) などを判定した。その結果をベースにさらに、上記文献集合 (1240) から 60 文献を抽出し、思考実験として自然語 (692 語) に対する統制語 (1048 語) との対応を詳細に解析し、実験システムの構築のための基本構想をまとめた。それは、次の 3 点である。

① 同形異義語の否定

基本的には語形が同一なら、意味は同一とする。

② 部分は全体の上位語

文字列 A が文字列 B に完全に含まれるならば、A は B の上位か関連概念とする。特に後方部に含まれる場合には A は B の上位概念となる例が多い。

③ 友達の輪

文字列 A の上位概念あるいは同義概念 A' を含む文字列 C が、文字列 B に含まれるならば、C は B の上位か関連概念である。

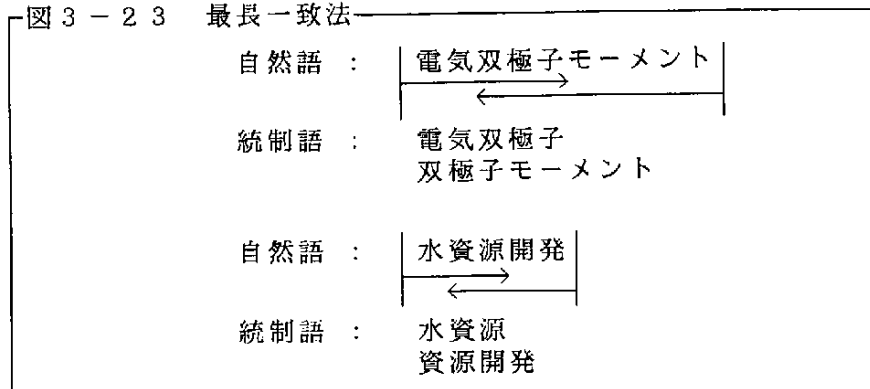
(c) 変換システムの概要・実験結果

本システムでは先頭および末尾よりの最長一致方式を採用した。(図 3-23) この抽出の精度および処理時間などの効率を上げるため、いくつかの処理を追加した。

① 前処理: 特殊な記号 (ハイフン、中黒) を含まぬ文字も統制語として抽出する。

② 抽出処理の終了: 抽出された統制語 (前後最大 2 語) の文字数合計が自然語をこえると

図 3 - 2 3 最長一致法



処理は終了する。

(例) 水資源開発 → 水資源 + 資源開発

③特殊処理：先頭および末尾が特殊文字については特定の統制語を発生させる。

(例) 耐凍害性 → 耐性 富士山 → 山岳

④最長一致した統制語に対して、置換辞書に登録されているその上位語、同義語、類義語と置き換えて新しい文字列（仮自然語）を生成し、再度最長一致方式で統制語を抽出し、先に抽出した統制語より長い文字ならばそれを採用する。

⑤文字列短縮：中間一致の代用として前方一致および後方一致した部分を削除した短縮文字列（仮自然語）を生成し、再度最長一致方式で統制語を抽出する。

⑥中抜き処理：中間にマスク文字の存在を許す「中抜き方式」による統制語の抽出も採用した。

以上の変換解析アルゴリズムをバッチ形式のプロトタイプシステムで検証した。(表3-13, 表3-14) その結果、化合物名、欧字綴りとカタカナ綴り音便(拗音、促音、長音)や音訳語などには課題は残るが、それなりの評価が得られた。

表 3 - 1 3 マニュアル索引との
照合一致度

(解析対象) :	5,794語
(完全一致) :	1601
(解析結果が余分)	1247
(不足は1/2以下)	678
(不足は1/2を越える)	101
(一致語なし)	1712
(対応CTなし)	455

表 3 - 1 4 実験に用いた辞書の概要

特殊処理	
特殊接辞々書72	接頭7 接尾65
置換処理(置換辞書)	
類義語尾	371
B T U S E / U F 置換	48074
照合(最長一致): 統制語抽出	
照合辞書(シソーラス)	48196

4. 辞書管理システムの概要

4.1. システムの設計

レコードマネジメント用辞書管理システムは、レコードマネジメントを目的に使用する索引語辞書（ファイル）に対し、新規の索引候補語を登録するシステムである。このシステムの設計にあたっての機能は図4-1のとおりである。

○単語入力方式

新規索引候補語に対し、概念分類を行い、単語分割を行い、概念分類体系画面内の概念名の項に入力する。

○辞書検索方式

入力新規索引候補の辞書内既蓄積索引語との形態照合処理を全辞書内既蓄積索引語対象完全一致と辞書内概念名後方一致にて行う。

○頻度評価

辞書検索の結果に対して不一致新規索引候補の頻度更新用一時蓄積単語との照合処理を行い、頻度値の判定処理を行う。

○頻度更新／一時蓄積

新規索引候補語として頻度ファイルの更新と一時蓄積を行う。

○辞書更新

新規索引候補語を索引語として同一概念下に登録する。

また、このシステムで用いられる辞書はISA関係構造シソーラス辞書である。ISA関係構造シソーラスのデータ構造は、基本的にはその分野における専門的概念（以下カテゴリという）のものとその分野で用いられる語（以下ディスクリプタという）を持っている。だが、ディスクリプタには表現が違っていても同じ意味や内容をもつものが多い。そのため、このシステムでは、同じ意味をもつ多くの候補語のうち代表となる用語をディスクリプタとして、それ以外の用語を同義語として登録する。

このシステムで用いるシソーラス辞書の概念的なデータ構造は図4-2のとおりである。

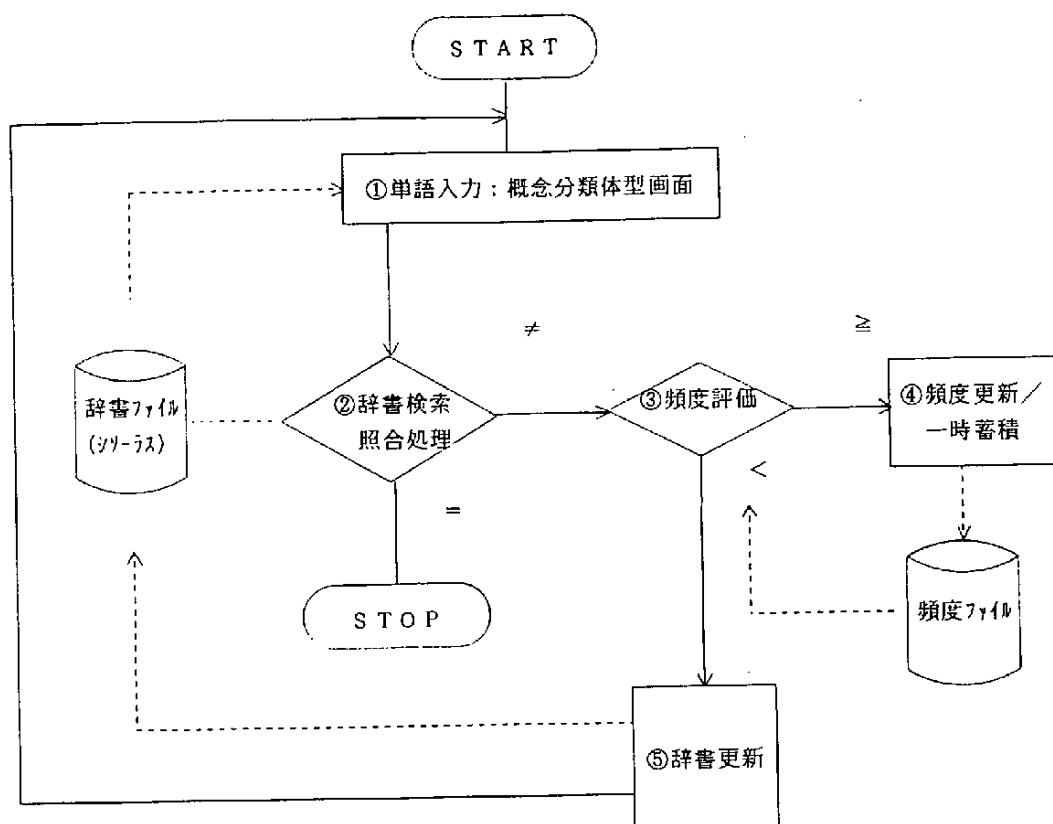


図 4 - 1 システム機能図

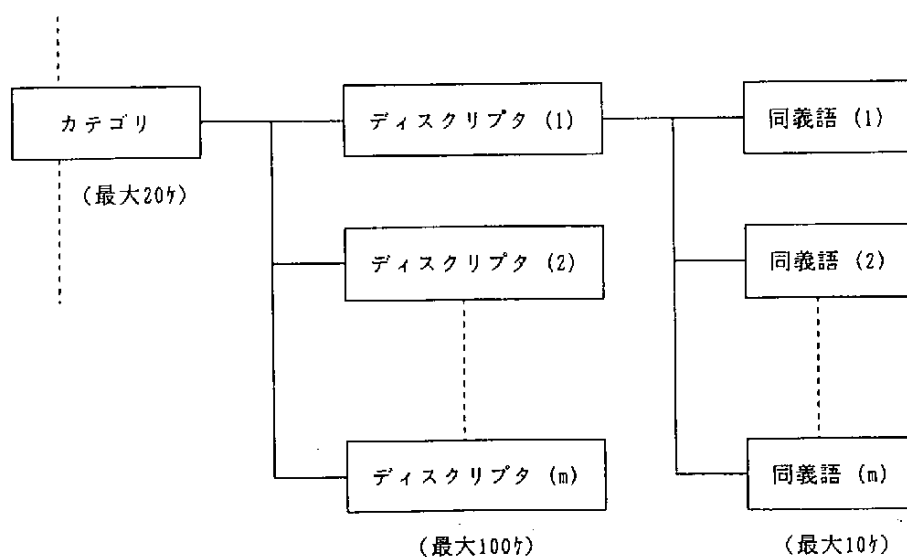


図 4 - 2 I S A ソーラスのデータ構造

今回のシステムの開発にあたって、辞書の容量算出のために当面目標とする登録語の数は、概ねカテゴリが20個、そしてひとつのカテゴリのもとにディスクリプタが100個、さらにそのディスクリプタのもとに同義語が10個程度を予定した。つまり、 $20 \times 100 \times 10 = 20,000$ 語程度が蓄積可能であるシソーラス辞書を想定した。この場合辞書の容量は、用語の平均文字数を経験な文字数であるが13文字程度とすると、520バイトが必要となる。これにインデックス部の容量を加味すると、1Mバイト程度つまり5インチのフロッピーディスク1枚分の容量である。

このシステムは一般の多くの会社、特に文字関係を扱う部所で利用されるものと思われる。分類されるカテゴリや登録されるディスクリプタの数については、その会社の業種や規模によっても違うものであろう。だが、1社無限にあるものではない。通常は、10数万語程度の用語を登録できる辞書を用意すれば充分であろうと思われる。以上のように複数の場所で利用され、その会社ごとに特徴あるシソーラス辞書となること、さらにその用語数の量などから、このシステムはパーソナルコンピュータ（以下パソコン）を用いて開発し、パソコン上にシステムを構築し、パソコンで操作できるシステムとした。パソコンの選定にあたっては、その普及率を考慮しPC-9800シリーズ（NEC社製）とした。また、OSについては同様の理由でMS-DOS（マイクロソフト社）を用いることにした。

さらに、システムの構築にあたっては、BASIC言語やC言語などに代表されるようなプログラム開発言語を用いる方法と、市販のデータベースマネジメントシステム（以下DBMS）を用いる方法について検討を行った。プログラム開発言語を用いる際の利点は、利用者指向の使い勝手のよいシステムが作成しやすいことにある。欠点は、例えばこのシステムにおけるシソーラス辞書のデータ構造は単純であるが、DBMSを作成しなければならないことである。そしてそのDBMSにはデータの蓄積、検索機能などDBMSとしてもたなければならないことである。そしてそのDBMSにはデータの蓄積、検索機能などDBMSとしてもたなければならない基本機能を作成しなければならない。また広く世の中に受け入れられるためには、データを蓄積するファイルの互換性も考慮しなければならない。その際には、ファイル形式の標準化の動向も加味しなければならないであろう。結果としてプログラム言語を用いて作成する方法は、時間と工数がかかることが予想される。

一方、市販のDBMSパッケージを利用する際の欠点は、そのDBMSパッケージの持っている言語の機能に依存してしまうため、利用者にとって使い易いと思われるようなシステムの実現は難しいことである。例えばマルチウィンドウでマウスを用いて、処理を選択するという機能が実現できなくなるであろう。だが、DBMSとしての機能が利用できるため、プログラム開発の面で作業工数が少なくなかつ短期間でできることが予想される。

以上のような検討内容から、今回のシステム開発には市販のDBMSパッケージを利用することとした。具体的には、dBASEⅢ（アシュトンテイト社）というパッケージを用いてシステムを開発する。その際、dBASEⅢのデータファイルと他に市販されているパッケージのデータファイルに互換性があることから、他のパッケージを利用できるという拡張性が生まれる。具体的には、このシステムで作成したデータファイルは市販のExcel（マイクロソフト社）やLotus（ロータス社）といった表計算パッケージでも読むことができるため、これらのパッケージを利用することは、結果的にExcelやLotusといったパッケージに用意されているすぐれたユーザインタフェースを利用することにもなる。

このシステムで扱うシソーラス辞書の用語は、カテゴリ、ディスクリプタ、同義語である。本システムでは図4-3のように、カテゴリ、ディスクリプタ、同義語それぞれに物理的に辞書（ファイル）をもち、その辞書間でカテゴリーディスクリプター同義語についてISA関係構造シソーラスとしての関係づけを論理的におこなう。つまり、本システムで扱うシソーラス辞書は、カテゴリ辞書、ディスクリプタ辞書、同義語辞書の3つから成り、それぞれが関係づけされている辞書である。

カテゴリ辞書のデータフォーマットは、図4-4のとおりである。カテゴリIDとは、カテゴリ登録時に連番で一意に付与されるもので長さとして5バイトで数字が与えられる。タームは、カテゴリの名称であり、最大38文字の領域が用意されている。ディスクリプタ数は、そのカテゴリの含まれるディスクリプタの個数であり、4バイトの数値データで表現される。候補とは、未だ正式なカテゴリとはなっていない用語、これをカテゴリ候補というが、その用語と正式にカテゴリとして登録されている用語との別を表現する。“F.”とある時はカテゴリ、“T.”とある時はカテゴリ候補である。

次にディスクリプタ辞書のデータフォーマットは図4-5に示すとおりである。ディス

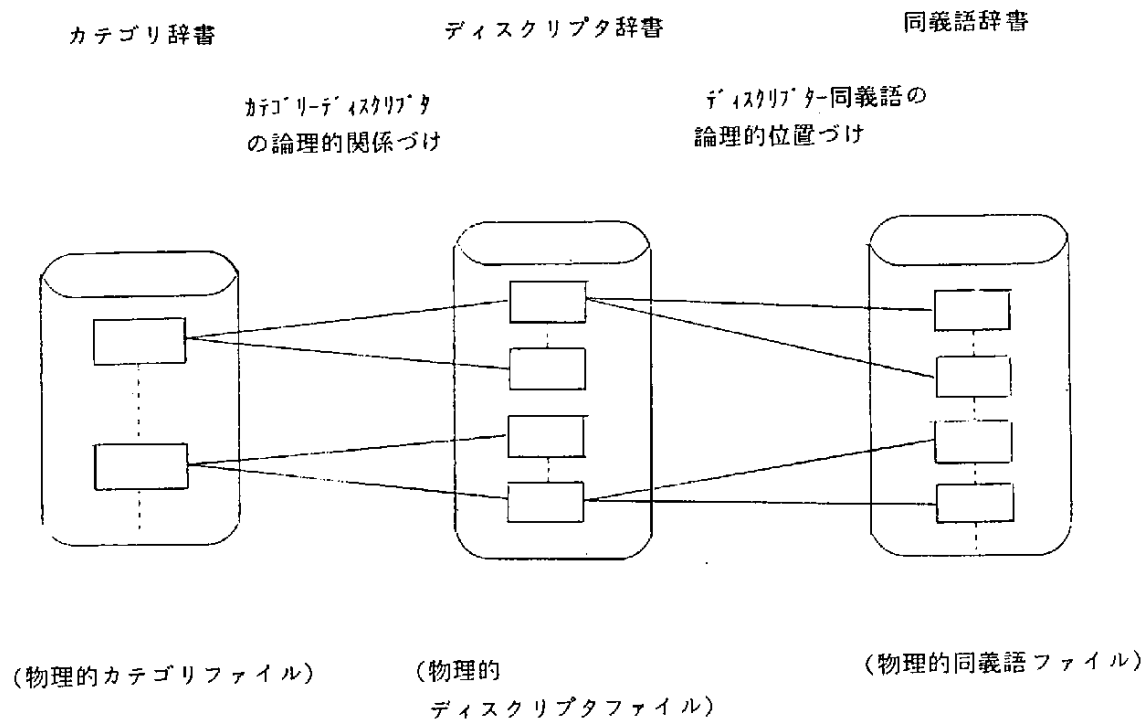


図 4 - 3 シソーラス辞書の構造

カテゴリ I D	タ ー ム	ディスクリプタ数	候 補
00001	事務用品	2	. F .
C(5)	C(16)	F(4)	C(1)

図 4 - 4 カテゴリファイルのデータフォーマット

クリプタIDとは、ディスクリプタ登録時に一意に付与される連番である。長さが5バイトの数字データより成る。タームは、ディスクリプタ名称そのもので最大38文字まで登録できる。逆タームは、タームのうしろの文字列より順に先頭文字列への配置となっており、後方一致検索の時に利用される。一致条件とは、ディスクリプタが登録される時、そのディスクリプタを含むカテゴリの配下にあるディスクリプタと後方一致したかどうかを判別するものであり、今回は後方一致が条件であることから‘B’という値が入っている。

一致文字数とは、後方一致となった時の後方からの文字数を表す。一致していない時は0が入る。同義語数は、そのディスクリプタのもっている同義語の個数を示している。上位カテゴリNo.とは、そのディスクリプタを含むカテゴリIDであり、ここでカテゴリとディスクリプタの論理的関係付けを行っている。最後のディスクリプタ出力順番号とは、ディスクリプタの表示および印字出力の際に利用するもので、上位にあるカテゴリのもとでこのディスクリプタが何番目に表示されるものであるかの情報をもっている。以上逆タームは38文字、一致条件は1文字、一致文字数、同義語数は4バイトの数値、上位カテゴリは5バイトの数字、およびディスクリプタ出力順番号は10バイトの数字で表現される。

同義語辞書のデータフォーマットは図4-6のとおりである。タームとは、同義語そのもので最大38文字の領域を用意している。上位ディスクリプタNo.とは、その同義語の代表としてのディスクリプタIDであり、これでディスクリプタと同義語との論理的関係付けを行っている。同義語出力順番号とは、表示および印字順を指定するもので、上位のディスクリプタのもとにこの同義語が出力される時、何番目に出力されるものであるかの情報である。上位ディスクリプタは5バイト、同義語出力順番号は10バイトの数字で表現される。

既にカテゴリ、ディスクリプタおよび同義語登録語数は総個数で20000語程度であることを記述したが、このシステムでは、その制限はもっていない。ただ各用語に付与される番号が数字の5バイトであるため、最大99999の番号をもつことが可能である。つまり、カテゴリの総数は最大99999、カテゴリの配下に付与されるディスクリプタの最大個数も99999個であり、ディスクリプタの下に付与される要義語の数も99999である。一方、ひとつのカテゴリに付与できるディスクリプタの論理的な最大個数は9999である。また、ひとつのディスクリプタに付与される同義語の数も同じである。以上の個数がこのシステムに

ディスクリプタ I D	ターム	逆ターム	一致 条件	一致 文字数	
00001	ノート	トーノ	B	0	
00002	ペ ン	ン ペ	B	0	
C(5)	C(76)	C(76)	C(1)	F(4)	

	同義語 数	上位 カテゴリNo	カテゴリ内の ディスクリプタ出力番号順
	1	00001	0000100001
	0	00001	0000100002
	F(4)	C(5)	C(10)

図 4 - 5 ディスクリプタファイルのデータフォーマット

ターム	上位ディスクリプタNo.	ディスクリプタ内同義語出力順番号
学 習 帳	00001	0000100001
C(76)	C(5)	C(10)

図 4 - 6 同義語ファイルのデータフォーマット

おける論理的な登録語数の最大である。

さらに、ディスクリプタの登録作業について、このシステムへ当初要求された機能を検討した結果、図4-7のような手順でディスクリプタの登録作業を行うことにした。その内容について簡単に説明する。

○ターム入力

当初このシステムへ要求された機能における単語入力に相当する。ディスクリプタの候補語（以下候補語）の入力には、利用者が直接画面に向かい、一語一語入力する方法と、フロッピーディスクから一括して大量の候補語に入力する方法があるが、本システムでは両方の入力方法を用意する。

○辞書との照合

要求機能における辞書検索に相当する。候補語と既に登録されているディスクリプタとの照合を行うもので、検索条件は完全一致にて行う。照合し一致した場合はディスクリプタ、同義語またはカテゴリとして既に登録済であるので、登録処理は行わない。不一致の場合は、頻度辞書との照合へ移る。なお、候補語との照合に用いる辞書はカテゴリ、ディスクリプタ、そして同義語辞書である。

○頻度辞書との照合

要求機能である頻度評価の辞書照合に相当する。候補語と頻度辞書との見出しマッチングを行う。完全一致した場合は、頻度数をカウントアップし「しきい値」（ディスクリプタとして登録する際の出語としての出現回数の最低値）との比較に移る。不一致の場合は、頻度数1の新しい用語として頻度辞書に登録を行う。そして「しきい値」との比較へ移る。

○しきい値との比較

要求機能である頻度評価の一部の機能と頻度更新／一時蓄積に相当する。候補語の出現回数と「しきい値」との比較を行う。候補語の出現回数が「しきい値」の値と等価になった時にシソーラス辞書の更新であるディスクリプタの登録作業を行うか、「しきい値」に達しない場合は頻度辞書の出現頻度数を更新する。

○シソーラス辞書の更新

要求機能における辞書更新に相当する。ディスクリプタ候補語を、どのカテゴリ配下のディスクリプタとして、またどのディスクリプタの同義語として登録するかの機能である。このシステムでは、登録しようとしている用語がどのカテゴリの配下に属するのかすでに

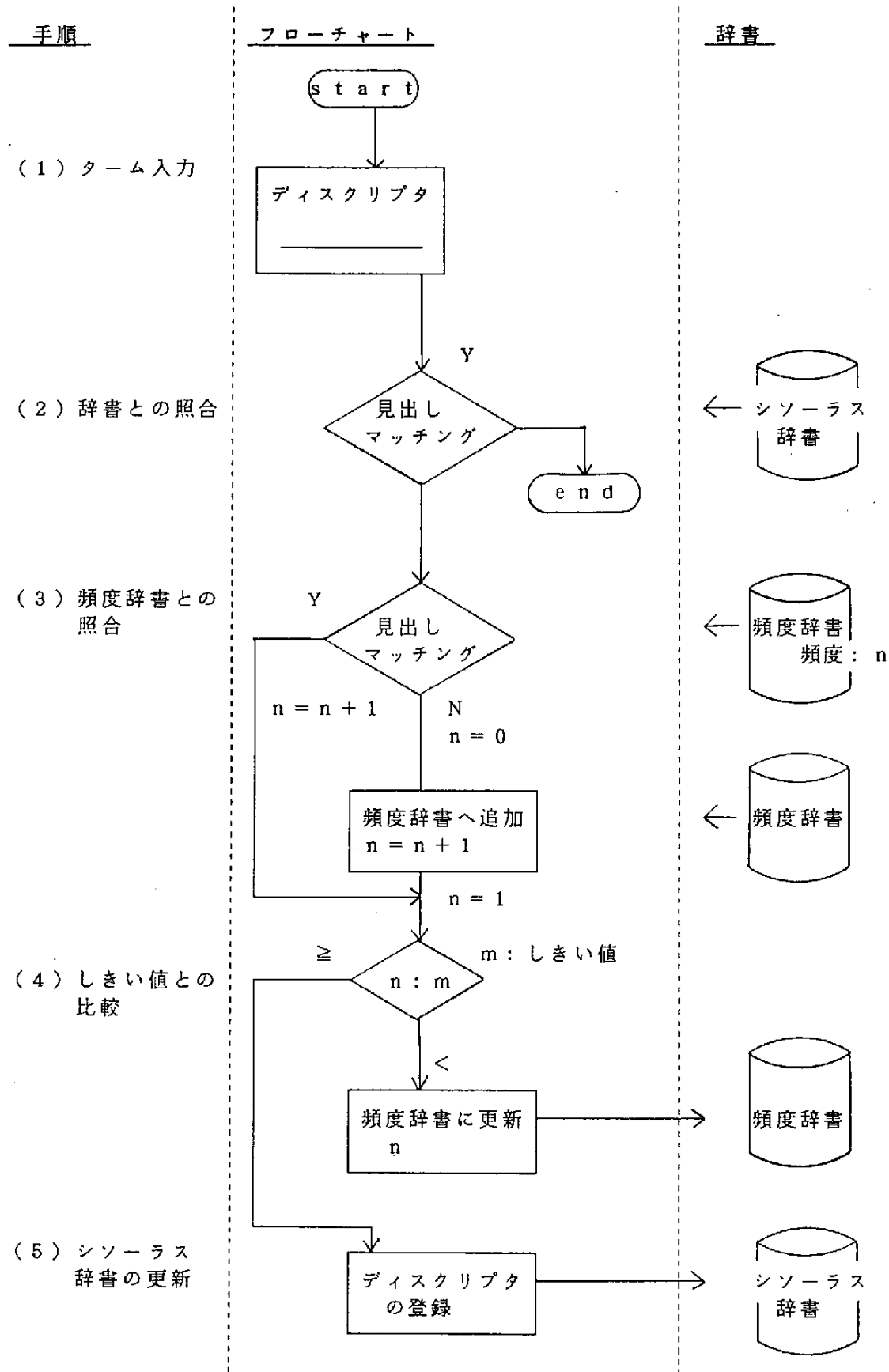


図4-7 システムの操作手順

登録されているディスクリプタを検索する。そして、既に登録されているディスクリプタと後方一致した時、検索されたディスクリプタを含むカテゴリが今登録しようとしている用語のカテゴリとして優先的であるとして利用者に通知する。利用者は、そのカテゴリを確認し、適切であればそのカテゴリのもとにディスクリプタとして登録する。検索条件は、末尾からの最長一致方式を採用した。例えば、カテゴリとして「会社名」で、その下に「岡村製作所」というディスクリプタが登録されており、候補語として「くろがね工作所」が入力されたとする。後方からの最長一致とは、最初「くろがね工作所」で検索を行い、該当する文字列をもったディスクリプタが存在しない時「くろがね工作所」の先頭から1文字ずつ減らしてマッチングパターンを作成し随時ディスクリプタを検索する方法である。この例では、「作所」で「岡村製作所」と「くろがね工作所」が一致する。この時「岡村製作所」を含むカテゴリである「会社名」を表示し、利用者にこのカテゴリのもとに「くろがね工作所」をディスクリプタとして登録してよいものかの確認をとれるようにしてある。なお、後ろの2文字でマッチングしない時つまり後ろ1文字でマッチングしてもこのシステムではそのカテゴリは採用されない。

次に、このシステムの機能について簡単に記述する。ディスクリプタの登録作業における○ターム入力、○辞書との照合、○頻度辞書との照合、○しきい値の比較という手順は、入力された候補語がディスクリプタとして登録するのに適切な用語であるが検査を行うものである。この間の手順はほとんど人間の判断を必要とせず、大量の候補語の中からディスクリプタとしての資格（ある「しきい値」をこえる）のある語を抽出するという一括（バッチ）処理的性格をもつ手続きである。一方、○シソーラス辞書の更新は、ディスクリプタとして登録資格のある用語がどのカテゴリのどのディスクリプタの同義語であるか判断しなければならないため、最終的には人間の知恵に頼らざるを得ないところである。以上のような理由からディスクリプタ登録作業のうち○ターム入力から○「しきい値」の比較までの処理を①「ディイクリプタ選択」としてひとつの手続きに、②シソーラス辞書の更新を②「ディスクリプタの登録」として2つの機能に分類した。

登録機能としては他に③「カテゴリの登録」に関する機能をもたせる。論理的には99999個までのカテゴリおよびカテゴリ候補語の登録を可能とする。カテゴリ候補語とは、今はその配下にディスクリプタが余りないがやや増加の傾向にあり、今後その分野の新しい専門的概念となる可能性が高い用語のことをいう。

次に用意するのは④「削除」の機能である。この機能はカテゴリ、ディスクリプタ同義語のいずれも削除可能とする。カテゴリを削除する場合はそのカテゴリに登録されているディスクリプタや同義語をそれぞれの辞書から削除しなければならない。カテゴリの削除は、カテゴリそのものが無くなってしまいうというより、カテゴリのもつ性格上時代とともに概念が変化し、例えばひとつのカテゴリが2つのカテゴリに分類されるとか、絶数のカテゴリをひとさのカテゴリに統一した方がふさわしいケースが生じるであろうことを考慮した。このため、カテゴリの削除によってディスクリプタ辞書および同義語辞書から削除される用語は、再度登録される可能性が非常に高いためこれらの語を一時的にファイルに蓄えることによって、ディスクリプタの登録機能を用いて登録機能を用いて再度登録できる方法を用意する。ディスクリプタや同義語の削除は誤操作や利用者の誤判断等によって、適正なカテゴリのもとにディスクリプタが登録されなかったり、適正なディスクリプタのもとに同義語が付与されなかった時に必要になるものである。同義語の削除は、同義語辞書より該当する見出し語を削除するとともに、ディスクリプタに登録されている同義語の個数を1減算する。ディスクリプタの削除は、該当する同義語があればその語のすべてを同義語辞書より削除した後、ディスクリプタ辞書より該当する見出し語を削除する。さらに、この語を含むカテゴリのディスクリプタ個数状況から1減算する。

この他に⑤「表示」と⑥「印刷」の機能を用意する。「表示」とはコンピューターのディスクリプタにカテゴリ、ディスクリプタや同義語を表示する機能である。ディスクリプタ画面は表示範囲に制限があるため1度にすべての用語に関する情報を表示するのは不可能である。そのためカテゴリ、ディスクリプタ、同義語というレベルで表示を行う。その際どのカテゴリの配下のディスクリプタに関して表示しているのか、さらにどのディスクリプタの配下の同義語を表示しているのか解るような表示方法をとる。「印刷」はシソーラス辞書として登録されているカテゴリ、ディスクリプタ、同義語を関連づけて一覧のリストを作成する機能である。

なお、このシステムではカテゴリやディスクリプタの文字列の修正機能を持っていない。理由は、あるカテゴリが複数に分類されるにしても統一されるにしてもカテゴリ名称のみの修正だけですむ問題ではなく、ディスクリプタに関する情報もあわせて検討しなければならないと思われる。さらに、ディスクリプタの修正については、登録されたディスクリプタの見出し語の誤りがその原因の多くであると思われるが、登録されたディスクリプタ

はある「しきい値」をこえるという検査に合格して登録されたものである。このような場合、用語の見出しに誤りがあっても先にその原因を明らかにすべきであろう。また、原因の多くがこの様なものであるなら、その出現頻度は非常に少ないものと思われる。以上のような理由から名称の修正は簡単に行うべきものではないように思われる。もし、そのような事態が発生した場合は、該当するカテゴリやディスクリプタを削除したのち再度キーボード入力によるディスクリプタの選択と登録機能およびカテゴリの登録機能を用いて追加する方法が残されている。

4.2. システムの操作

レコードマネジメント用辞書管理システムは、市販のDBSパッケージであるdBASE IIIを用いて作成している。このパッケージのもっている機能で作成できる画面の表示方式や操作性などに代表されるユーザインタフェースは、必ずしもよいと言えない。だが、このシステムの利用者の多くは、情報処理という分野では素人と言われる人々であることが想定される。そのため、このシステムの作成時にあたっては、特に簡便な操作で利用できるシステムとなるように注意し設計と開発を行った。

具体的には、ウィンドウ内に表示されているメニューの選択などは、目的とするメニューに移動する操作は「カーソルキー」、目的としたメニューの選定は「リターンキー」という2つのキー操作のみで行えるようにしてある。また、利用者がこのシステムのどこにいて、何をしようとしているかわからなくなった時などは、構造化されているシステムのどこにいても、「ESCキー」を押すという行為のみでシステムの初期のメニューに戻れるようにしてある。さらに、ディスクリプタやカテゴリの登録、および削除といった特別の注意を要する機能は、キーの配置位置としてキーボードの上位に位置し、通常のキーボードキーの位置すら少し離れているファンクションキーに割り当てることによって、ユーザの行為の確認を行えるようにし、操作上の不安をなるべく軽減するよう配慮してある。

このシステムは、dBASE IIIで作成していることは前述したが、このシステムを立ち上げる前に、MS-DOSのもとでまでDBASE(dbase)というコマンドを入れ、dBASE IIIのシステムを立ち上げておかなければならない。その後「DO MEN

U」と入れることによって、レコードマネジメント用辞書管理システムのプログラムが起動される。

このシステムを操作するにあたり、利用者は次のような点に注意をして欲しい。まず、作成してあるプログラムは、インタプリタといわれるその都度プログラムを解析しながら実行するモードで走行するために、実行速度が遅く、利用者の行為へのリアクションが若干遅くなる。そのため「ESCキー」などを多用すると、このシステムではなくdBASEⅢのシステムへ制御がわたってしまい、dBASEⅢのシステムが「ESCキー」をひろってしまうことがある。こんな時は、「I」のIgnore（無視）を選択し、一担このシステムへ制御を戻し、ゆっくりと「ESCキー」を押すことによって、このシステムの初期の画面まで戻ることが可能である。

このシステムが起動されると図4-8の初期画面が表示される。この初期画面には、レコードマネジメント用辞書管理システムの持っている機能のすべてが見られるようにしてある。つまり、利用者は表示されたメニューの内容を見ることによって、このシステムでは何ができるかを知ることができる。

以下簡単に各機能の操作について、画面を示しながら説明する。

4.2.1 候補語の選択

初期画面より候補語の選択を行うと、図4-9にある入力方式の選択画面が表示される。このメニューではフロッピーディスクから一括して候補語を入力する方法をとるか、その場でコンピュータと会話しながら候補語を入力する方法のどちらかにするかの選択を行う。

○フロッピーディスクによる一括入力

フロッピーディスクによる一括入力方法を選択すると、このシステムで持っている「しきい値」が表示される。変更したければその値を入れる。すると、図4-10にあるフロッピーのドライブを入れるデバイスを指定する画面が表示され、指定されたドライブから一括入力方式による候補語の選択処理が行われる。もし、指定されたドライブに該当するファイル名が存在しない時は、図4-11のようなエラーメッセージが出力される。

フロッピーディスクによる一括入力時の制約事項として、一括入力をするファイル名を固定にしてあることである。そのファイル名は、「TERMDB.TXT」である。そのためこのシステムで候補語を処理するためには、この名前で作成しておくか、またはMS-DOSの「renameコマンド」で名前を変更しておく必要がある。

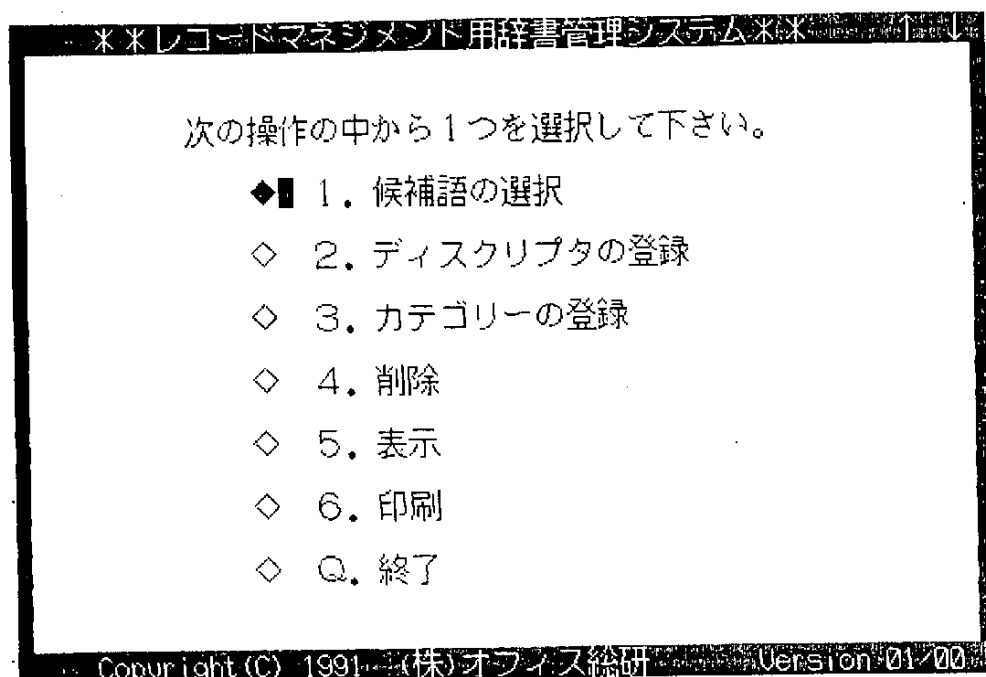


図4-8 初期化画面と候補語の選択

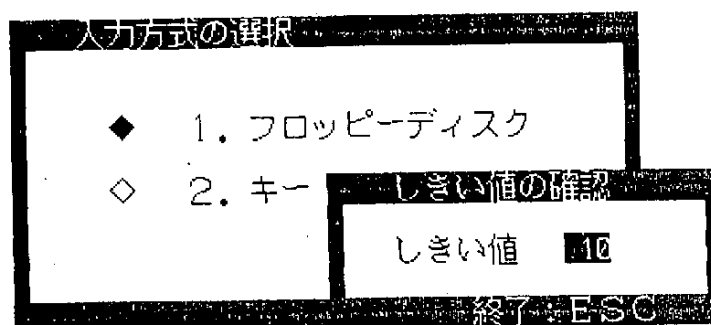
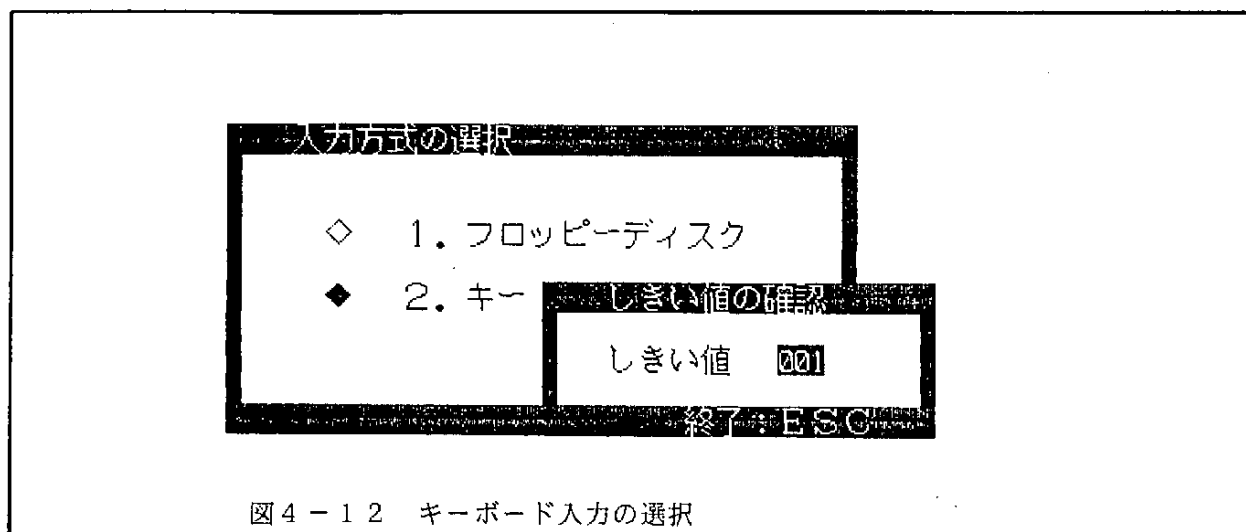
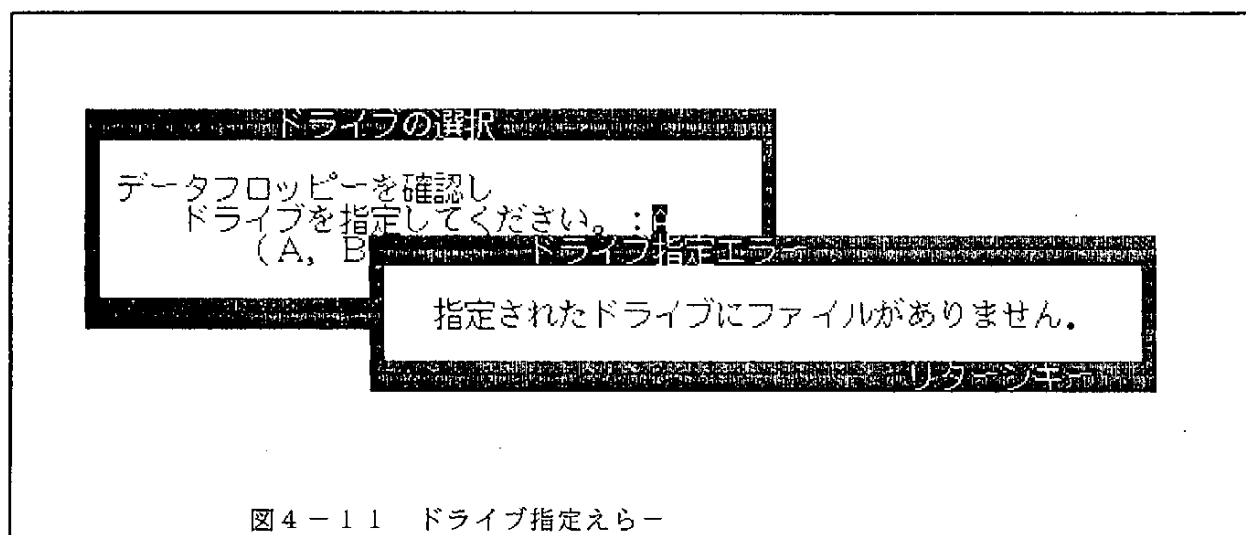
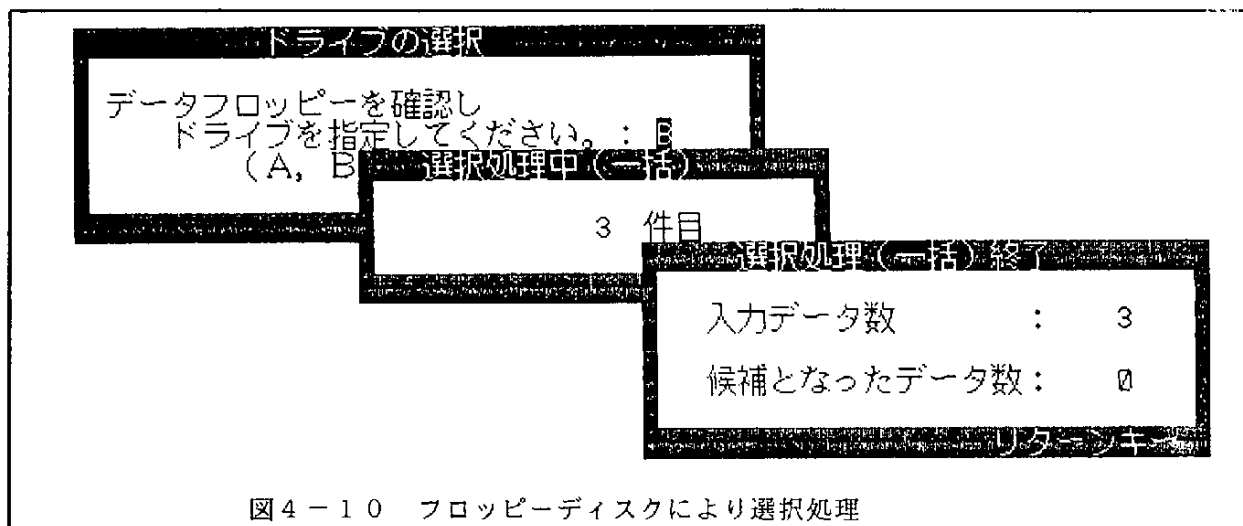


図4-9 フロッピーディスク入力の選択



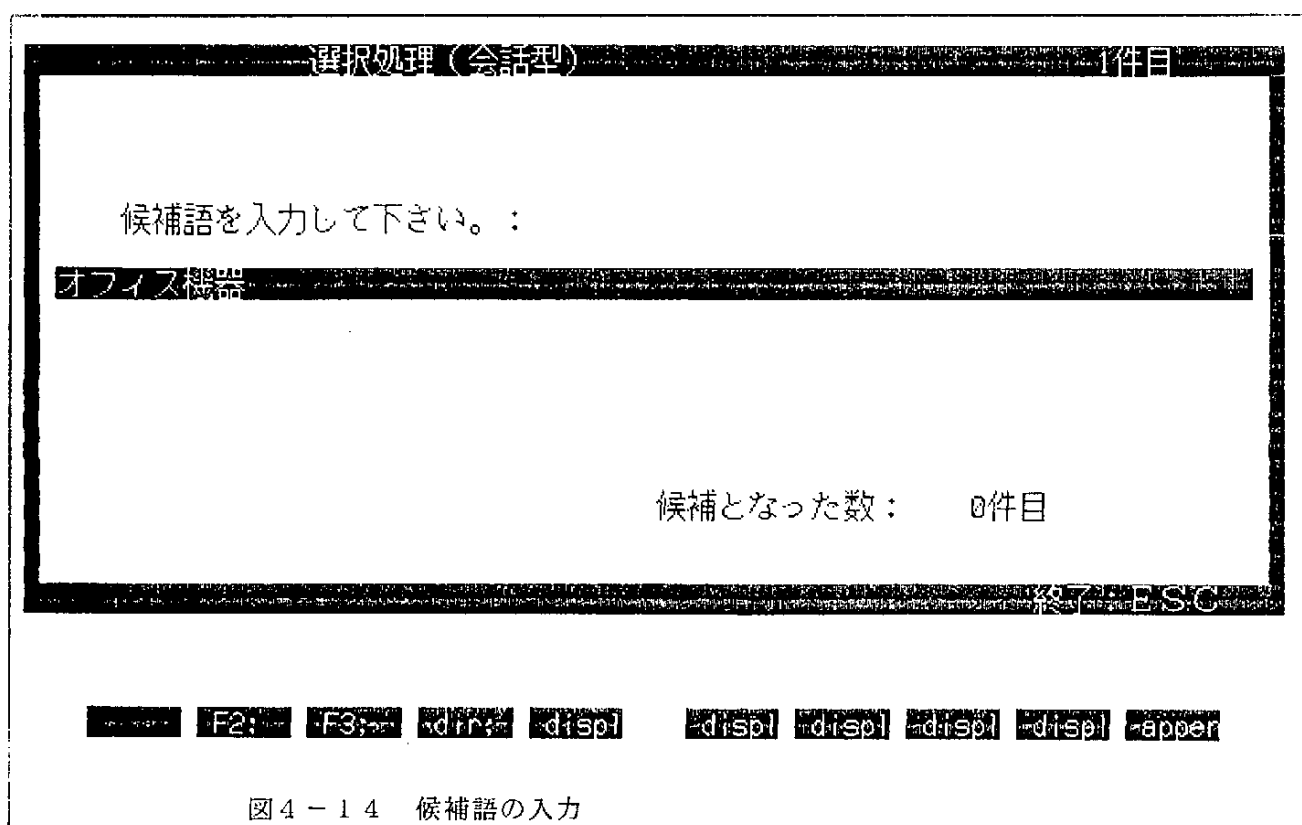
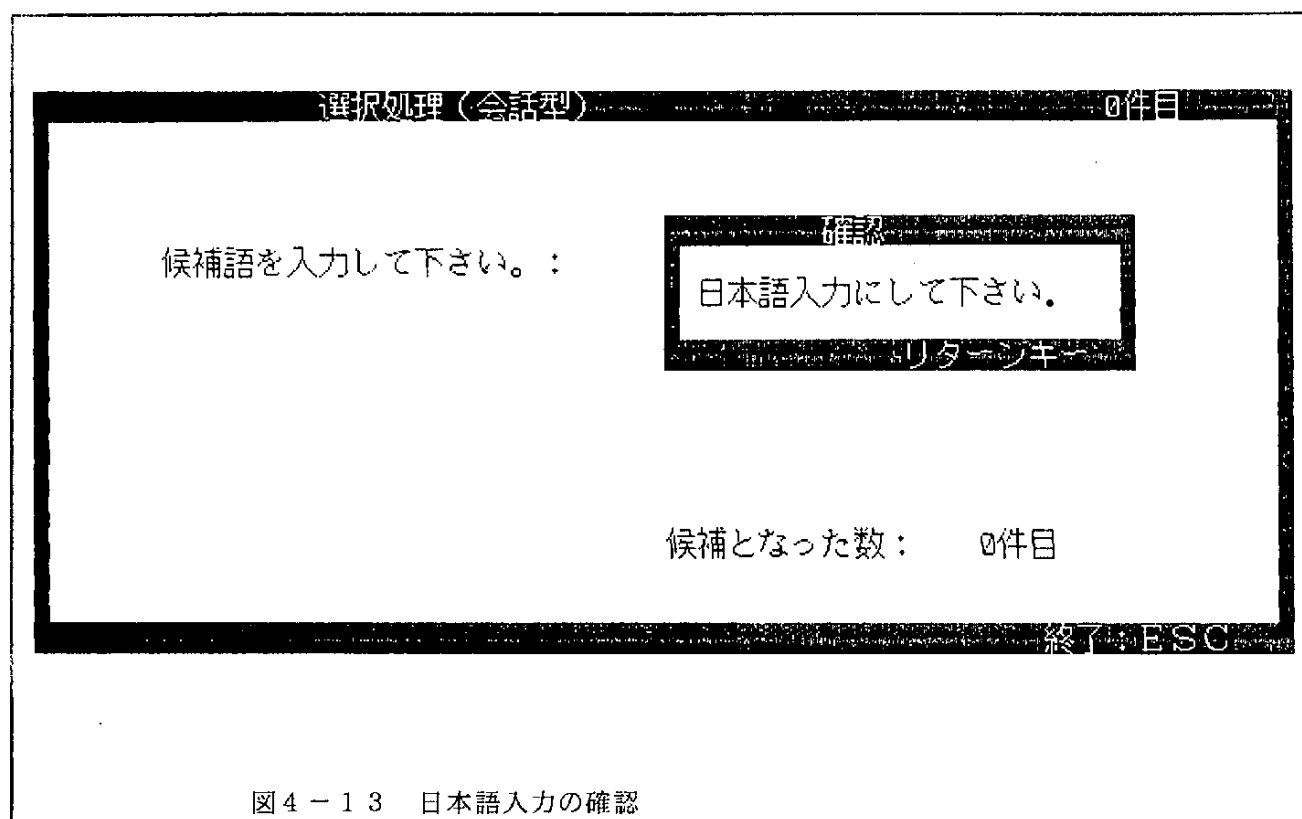
○キーボードによる入力

図4-12のメニューでコンピュータと会話しながら候補語をキーボードから入力する方法を選択すると、フロッピーから一括して入力する方法と同じく「しきい値」の確認が行われる。そのあと、図4-13で日本語データの入力モードになっているかどうかの確認画面が表示され「リターンキー」を押すことによって図4-14にみるような候補語の入力画面になる。この時点で、ウィンドウの枠にある入力件数が1件目となる。ここで候補語となる語を入力すると、シソーラス辞書との完全一致の照合を行い、シソーラス辞書にないものをディスクリプタに登録する候補語とする。そして、次の候補語の入力促進となる。その画面が図4-15である。ウィンドウの枠の右上に表示してある入力データ数が2件目の入力であることを示し、入力された候補語のうちディスクリプタの候補語として1件採用されていることを枠内に表示している。

4.2.2 ディスクリプタの登録

図4-16の初期画面よりディスクリプタの登録を選択する。図4-17にみる入力ファイルの選択画面が表示される。候補語ファイルとは、候補語の選択処理によってディスクリプタ候補語となった語が入っているファイルである。削除語ファイルとは、カテゴリが削除されたときに作成されるファイルで、削除されたカテゴリ配下のディスクリプタと同義語が格納されている。どちらかを選択したとしても、そのファイルに何も候補語が入っていなければ、登録処理は終了し、図4-18のようなエラーメッセージが出力される。

候補語が入っている場合は、文字列の後方一致検索を行う。後方一致となる語がディスクリプタおよび同義語に存在した場合は、図4-19のようにその語を含むカテゴリが表示される。例では、「オフィス機器」というカテゴリしか表示されていないが、もし他にあればそのカテゴリ名が表示されるし、「オフィス機器」の中のディスクリプタレベルで複数一致した場合は、一致した語の数だけ「オフィス機器」という同じカテゴリ名が表示される。この例では、ディスクリプタ候補語のウィンドウに、今処理対象となっている語が表示されているが、これによってどのような用語を処理しているかがわかるようにしている。このカテゴリの表示では、まだ登録処理は行うことはできない。登録処理を行うためには、この表示されているカテゴリの中から登録したいカテゴリを選択し、ディスクリプタの登録画面または同義語の登録画面にすることによって可能となるが、詳しくは後述



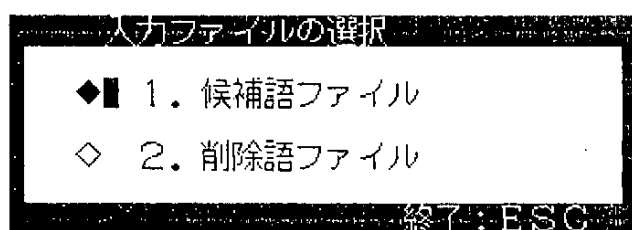


図 4 - 1 7 入力ファイル選択画面



図 4 - 1 8 候補語または削除語ファイルが空の時

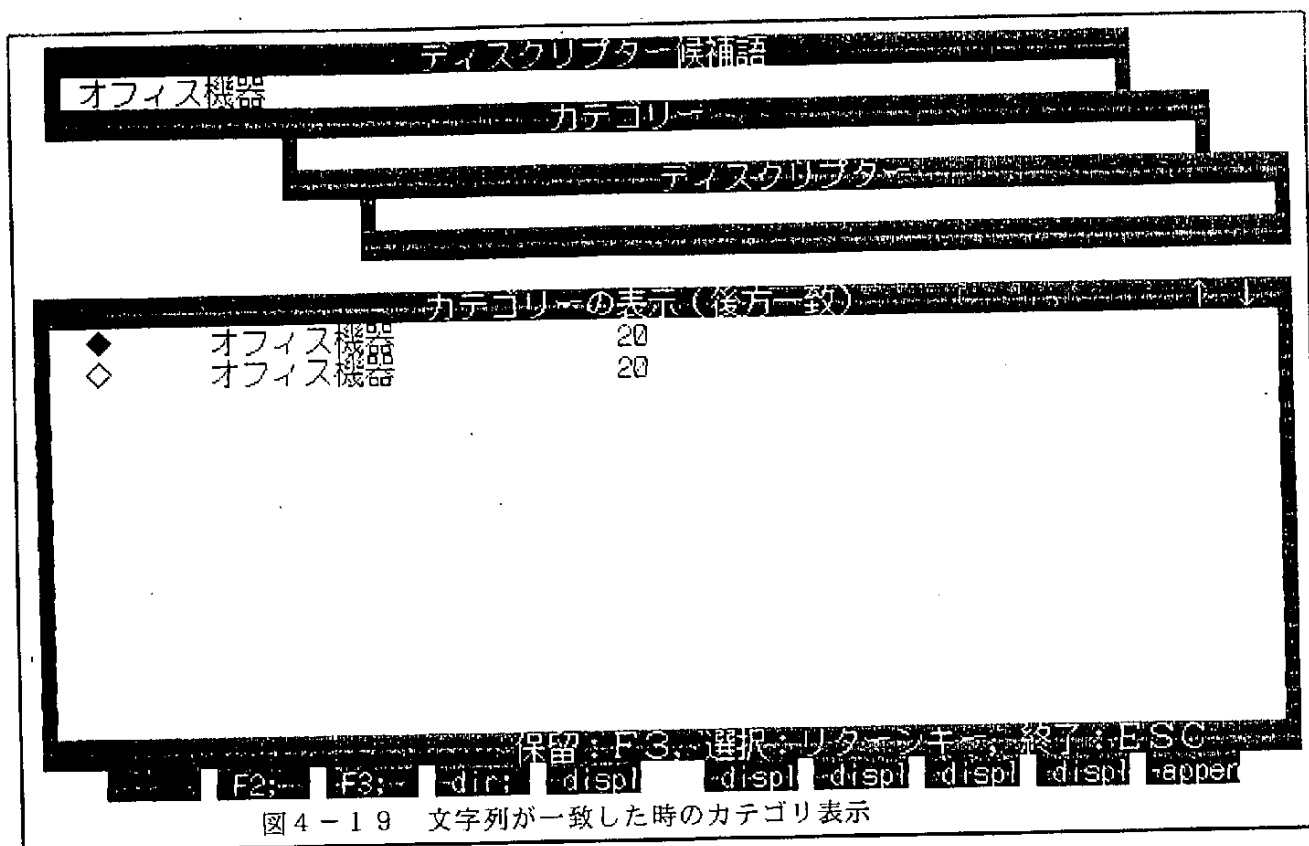


図 4 - 1 9 文字列が一致した時のカテゴリ表示

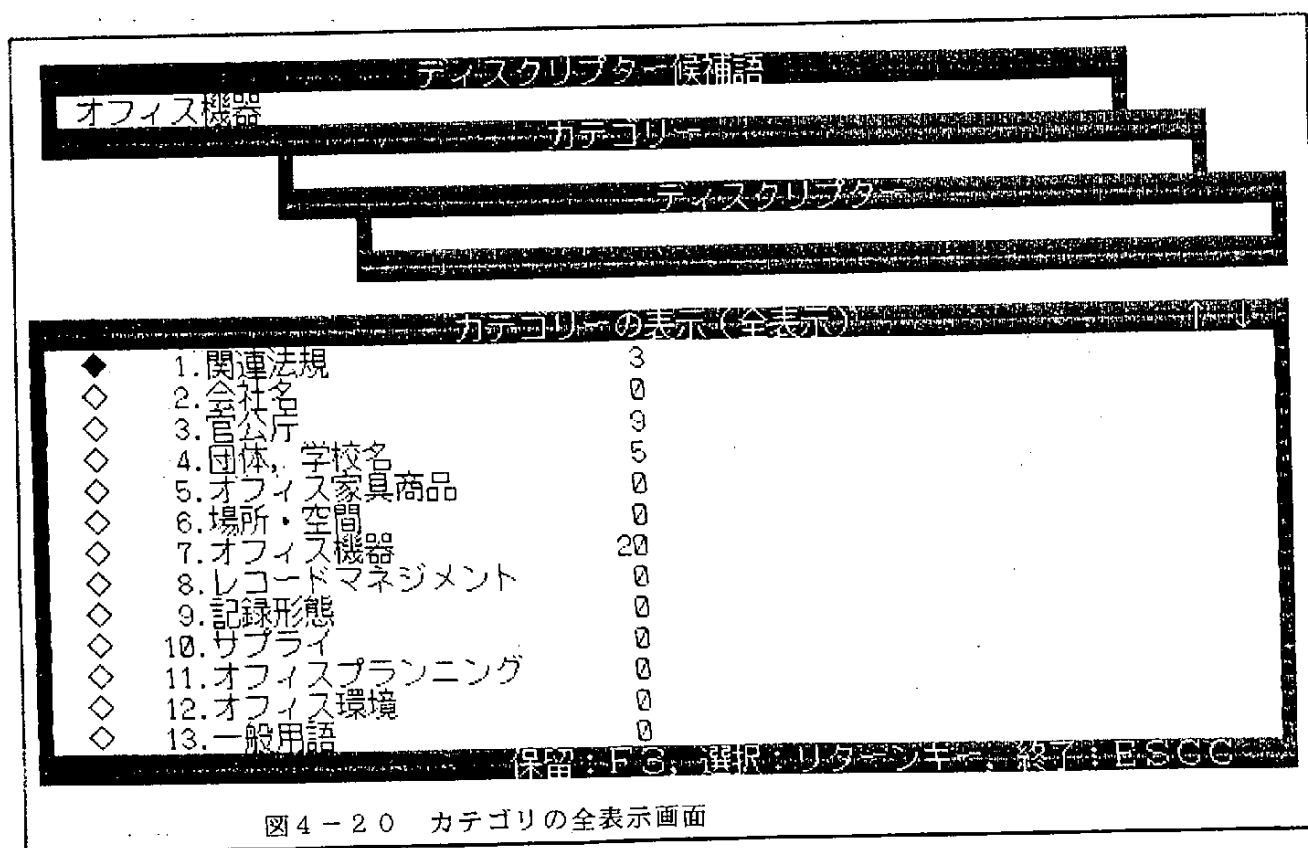


図 4 - 2 0 カテゴリの全表示画面

する。もし、ここでこの候補語を登録するのに該当するカテゴリがないため他のすべてのカテゴリを見たい場合は、「ESCキー」を押すことによって、図4-20のカテゴリの全表示へと移ることができる。また、該当する候補語の処理を中断し、次の候補語の処理を行いたい時は、「保留：F3」を押すことによって次の候補語による検索が行われる。

図4-20は、「オフィス機器」という候補語によってシソーラス辞書を検索し、後方による一致をみななかった場合の例である。図4-20のカテゴリの全表示ウィンドウにおいては、2列13行のカテゴリ、つまりひとつのウィンドウに最大26ヶのカテゴリが表示可能（これはディスクリプタでも同義語でも同じ）である。カテゴリ表示において、表示されているカテゴリ候補（まだカテゴリに昇格していない語）であった場合、表示列の一番先頭に‘@’表示される。その右の‘◇’の列の‘◆’は、現在システムとして注目しているカテゴリ、その右の数は表示している順番号、そしてカテゴリ名、それからそのカテゴリが保有しているディスクリプタの数（ディスクリプタが表示されている場合は、そのディスクリプタのもつ同義語の数）が表示される。例として、オフィス機器というカテゴリを選んだとする。そうすると、図4-21のような候補語をディスクリプタとして登録できる画面となる。この時、カテゴリのウィンドウには、表示されているディスクリプタが含まれるカテゴリ名が表示される。さらに、1.「ワープロ」というディスクリプタを選ぶと図4-22のように「ワープロ」というディスクリプタのもつ同義語が表示される。ここで同義語の登録が可能となる。

ここでは例として、オフィス機器の下に「オフィス機器」というディスクリプタを登録してみる。図4-23のディスクリプタウィンドウで、18.の「端末」のあとに「オフィス機器」というディスクリプタを登録したいとする。その時はまず「カーソルキー」を18.の「端末」の位置まで持っていく、「F2登録キー」を押す。これで「オフィス機器」は「オフィス機器」というカテゴリのディスクリプタとして登録される。この画面例では、ディスクリプタのウィンドウ以外の各ウィンドウにデータが表示されている。まず、候補語ウィンドウでは「オフィス機器」という語が処理中であり、今表示しているディスクリプタは、「オフィス機器」というカテゴリの配下のディスクリプタであることが解るようになっていく。このような操作を繰り返すことによって、候補語ファイル又は削除語ファイルにある語の登録が可能である。これらのファイルにあるすべての語の処理が終了すると図4-24の登録処理の終了のウィンドウでその件数が通知される。その後、表示の機能

ディスクリプター候補語

オフィス機器

カテゴリー

オフィス機器

ディスクリプター

ディスクリプターの登録

◆	1. ワープロ	2	◇	14. VDT	0
◇	2. パソコン	1	◇	15. ゲートウェイ	0
◇	3. ラップトップ	0	◇	16. コンピュータ	1
◇	4. 光ファイル装置	0	◇	17. ホストマシン	0
◇	5. マイクロフィルムリーダー	0	◇	18. 端末	0
◇	6. キーボード	0	◇	19. 照明機器	0
◇	7. マウス	0	◇	20. 空調機器	0
◇	8. プリンター	0			
◇	9. レーザープリンター	0			
◇	10. ドットプリンター	0			
◇	11. CVCF	1			
◇	12. CAD	0			
◇	13. CRT	0			

登録: F2, 保留: F3, 選択: リターンキー, 終了: ESC

図4-21 ディスクリプターの登録画面

ディスクリプター候補語

オフィス機器

カテゴリー

オフィス機器

ディスクリプター

ワープロ

同義語の登録

◆	1. W/P
◇	2. WP

登録: F2, 保留: F3, 選択: リターンキー, 終了: ESC

図4-22 同義語の登録画面

オフィス機器

カテゴリー

ディスクリプター

ディスクリプターの表示

◆	1. ワープロ	2	◇	14. VDT	0
◇	2. パソコン	1	◇	15. ゲートウェイ	0
◇	3. ラップトップ	0	◇	16. コンピュータ	1
◇	4. 光ファイル装置	0	◇	17. ホストマシン	0
◇	5. マイクロフィルムリーダー	0	◇	18. 端末	0
◇	6. キーボード	0	◇	19. オフィス機器	0
◇	7. マウス	0	◇	20. 照明機器	0
◇	8. プリンター	0	◇	21. 空調機器	0
◇	9. レーザープリンター	0			
◇	10. ドットプリンター	0			
◇	11. CVCF	1			
◇	12. CAD	0			
◇	13. CRT	0			

選択: リターンキー, 終了: ESC

図 4 - 2 5 ディスクリプタの表示

レコードマネジメント用辞書管理システム

次の操作の中から1つを選択して下さい。

- ◇ 1. 候補語の選択
- ◇ 2. ディスクリプタの登録
- ◆ 3. カテゴリーの登録
- ◇ 4. 削除
- ◇ 5. 表示
- ◇ 6. 印刷
- ◇ Q. 終了

Copyright (C) 1991 (株) オフィス総研 Version 01/00

図 4 - 2 6 カテゴリー登録時の初期画面

を用いて「オフィス機能」というカテゴリ配下のディスクリプタをみたものが図4-25である。候補語のあとに「オフィス機器」というディスクリプタが表示されている。

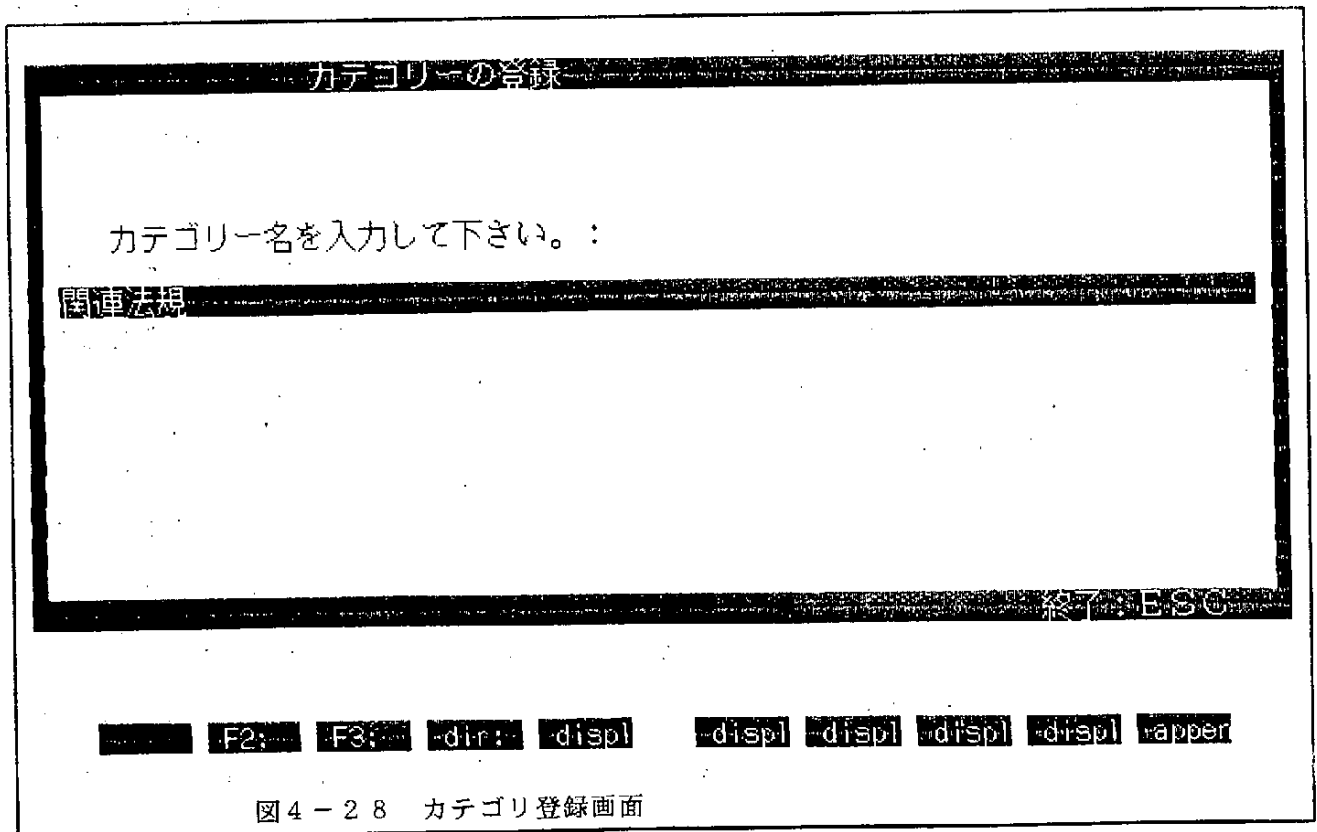
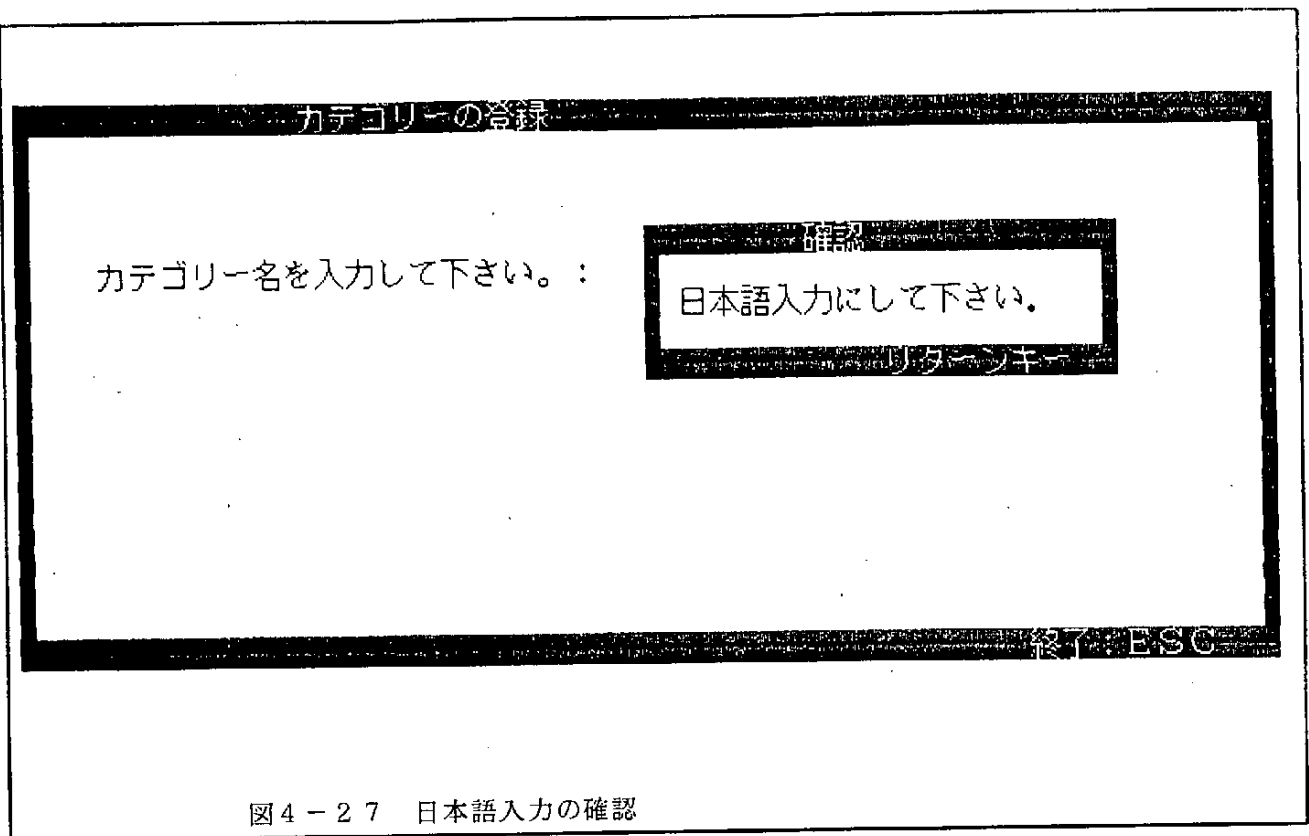
4.2.3 カテゴリの登録

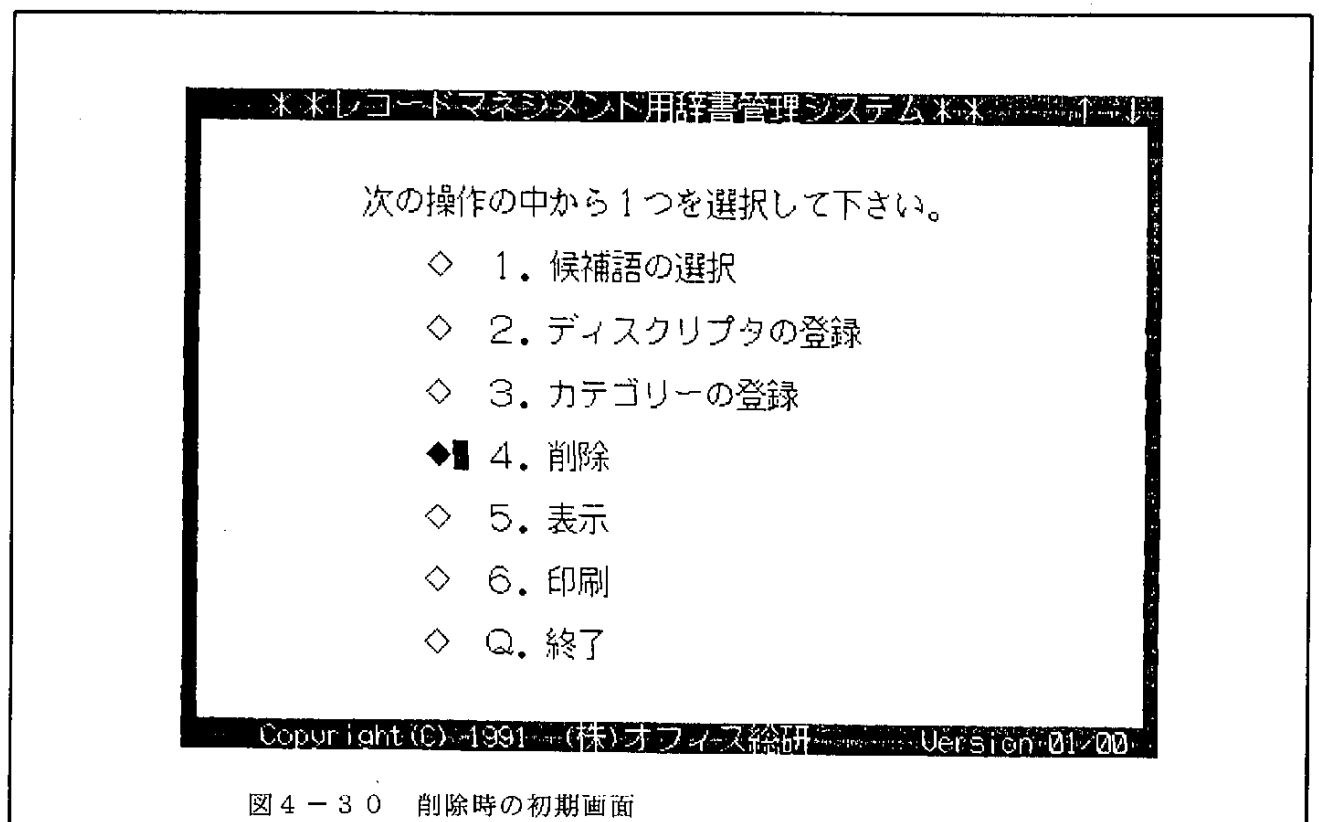
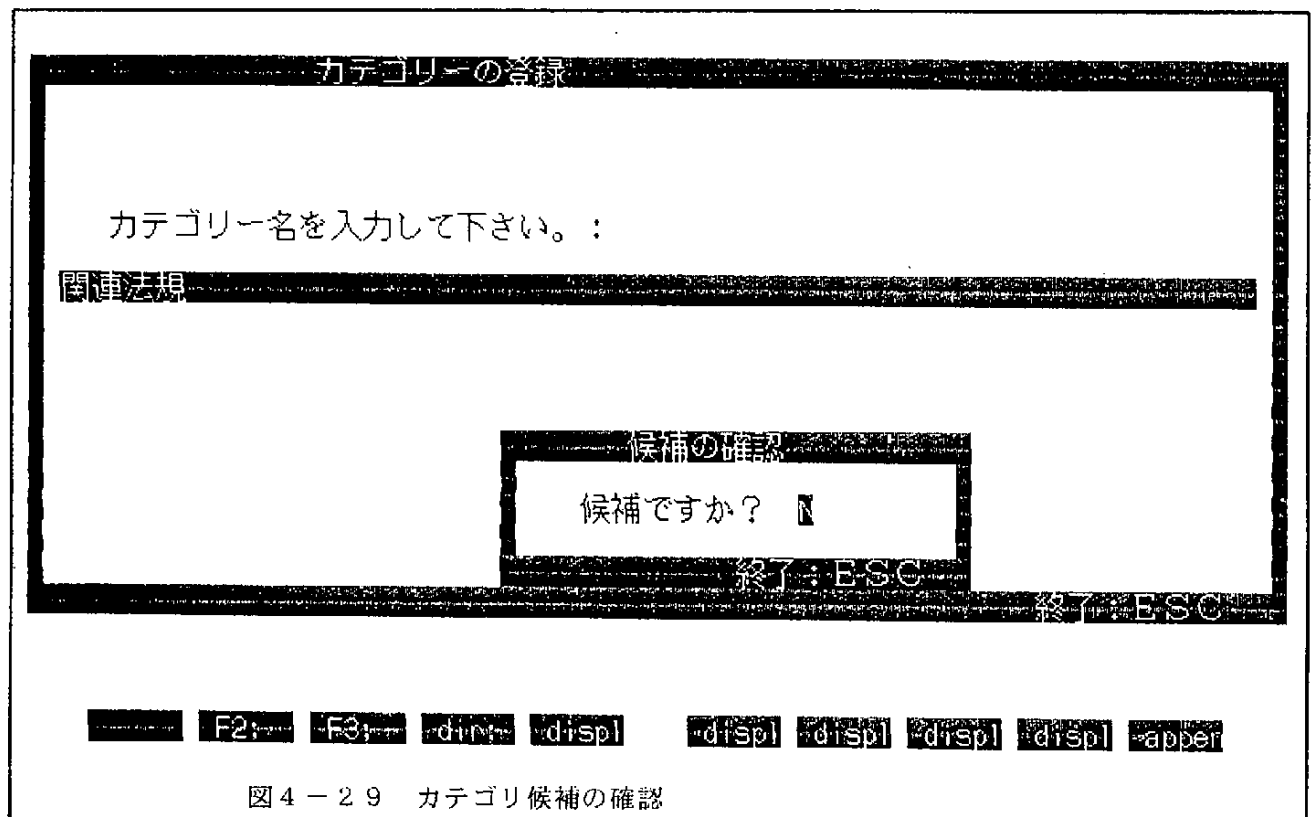
操作はキーボード入力による会話型の候補語選択処理に似ている。図4-26の初期画面でカテゴリ登録を選択すると、図4-27にみる日本語の入力モードの確認がされる。

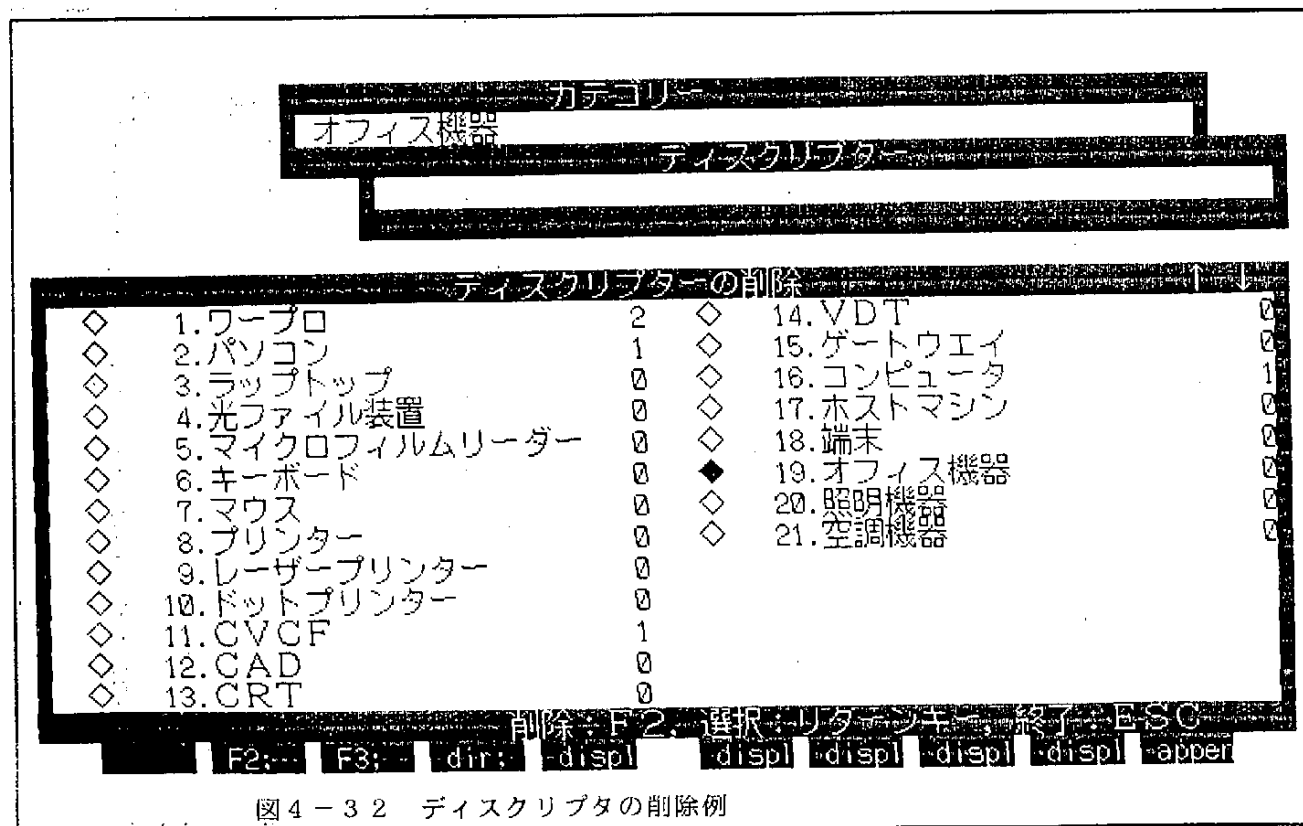
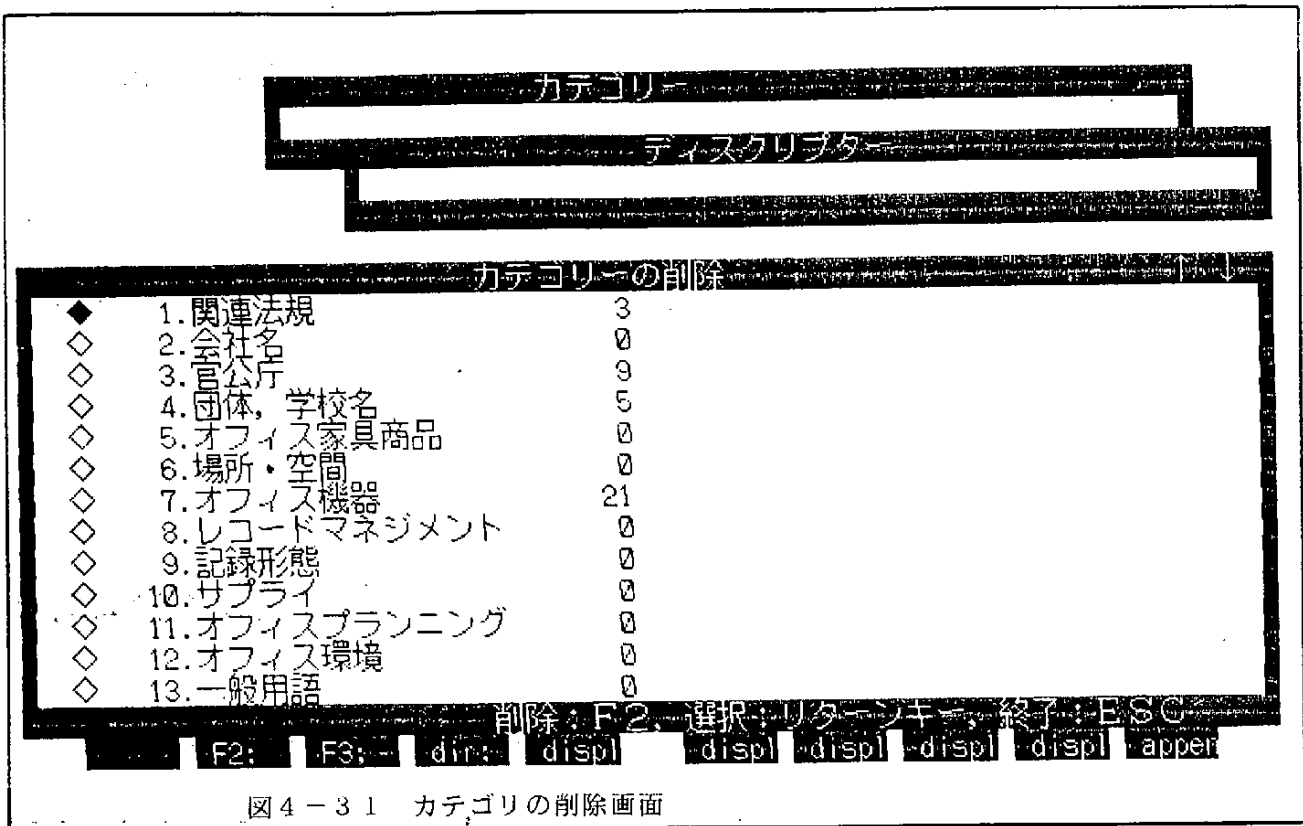
「リターンキー」を押すことによって図4-28の画面に移りカテゴリ登録が可能となる。ここで登録したいカテゴリを入力すると、図4-29にみるカテゴリ候補かの確認画面が表示される。カテゴリなら「リターンキー」、カテゴリ候補として登録するなら「Y」と入れて「リターンキー」を押す。すると次に登録するカテゴリの入力促進面となるが、終了したければ「ESCキー」で初期画面へと戻れる。

4.2.4 削除

この削除は、カテゴリ、ディスクリプタ、同義語すべてを削除することが可能である。図4-30の初期画面で削除機能を選択する。すると、図4-31のカテゴリ削除画面が表示される。この例でみるように、カーソルを「オフィス機器」というカテゴリの位置に移動させ「リターンキー」を押したりすると、図4-32のようにカテゴリウィンドウに「オフィス機器」が表示され、ディスクリプタ削除ウィンドウに「オフィス機器」の配下のディスクリプタが表示される。さらに、カーソルを「ワープロ」というディスクリプタの位置に移動させ「リターンキー」を押すと、図4-33のようにカテゴリが「オフィス機器」でディスクリプタが「ワープロ」の同義語が表示される。削除する際は、削除したいカテゴリ名、ディスクリプタ名、同義語名の位置にカーソルを移動させ、「削除キーであるF2」を押すことによって可能である。図4-34は、「オフィス機器」というディスクリプタを削除している例である。まず、「オフィス機器」の位置にカーソルを持っていき「F2」を押すと、削除対象語のウィンドウがオープンされ削除対象語である「オフィス機器」が表示されると同時に、確認のメッセージウィンドウがオープンされる。削除する語であれば「リターンキー」を押すことによってその語は削除される。もし、誤操作であるなら「N」を入れてリターンキーを押すことによってその語は削除されない。「オフィス機器」を削除したのちのディスクリプタをみたものが図4-35であり、「オフィ







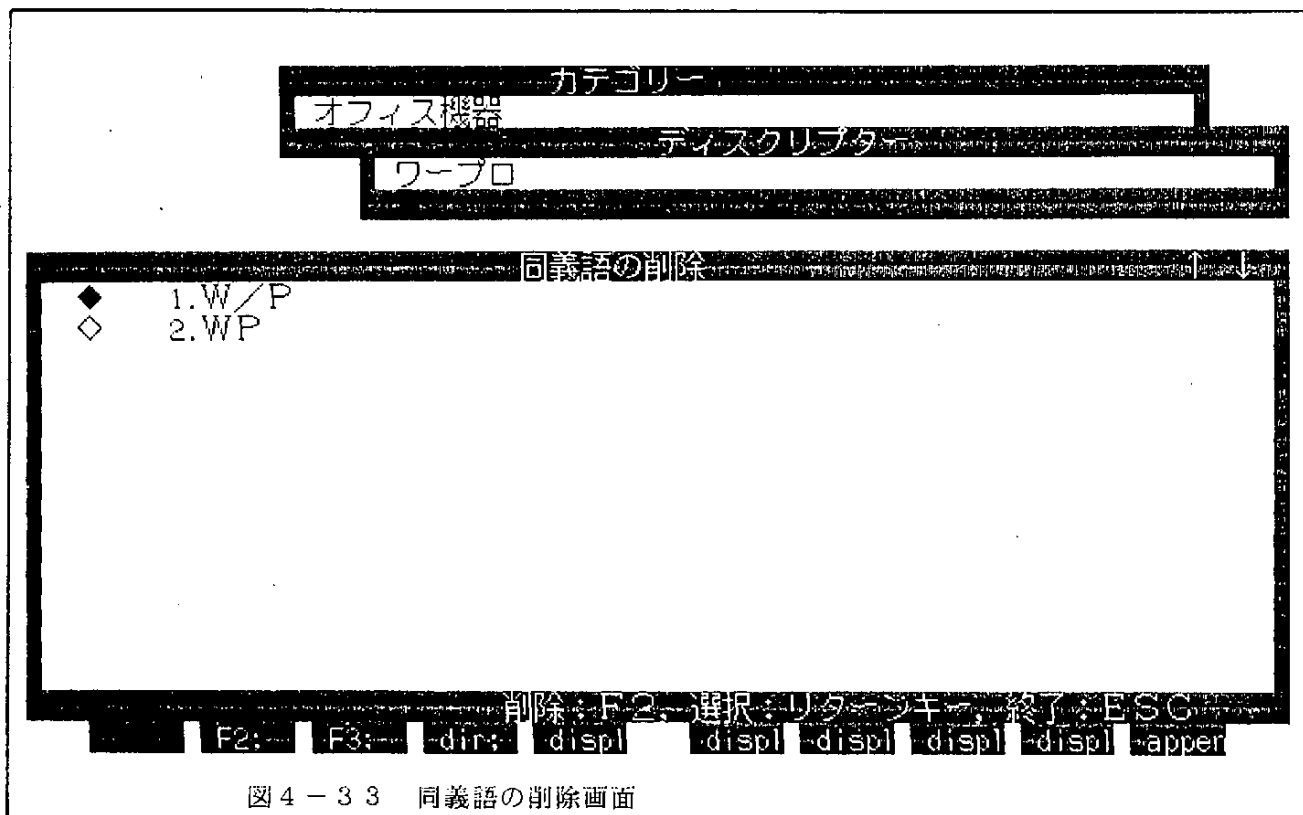


図4-33 同義語の削除画面

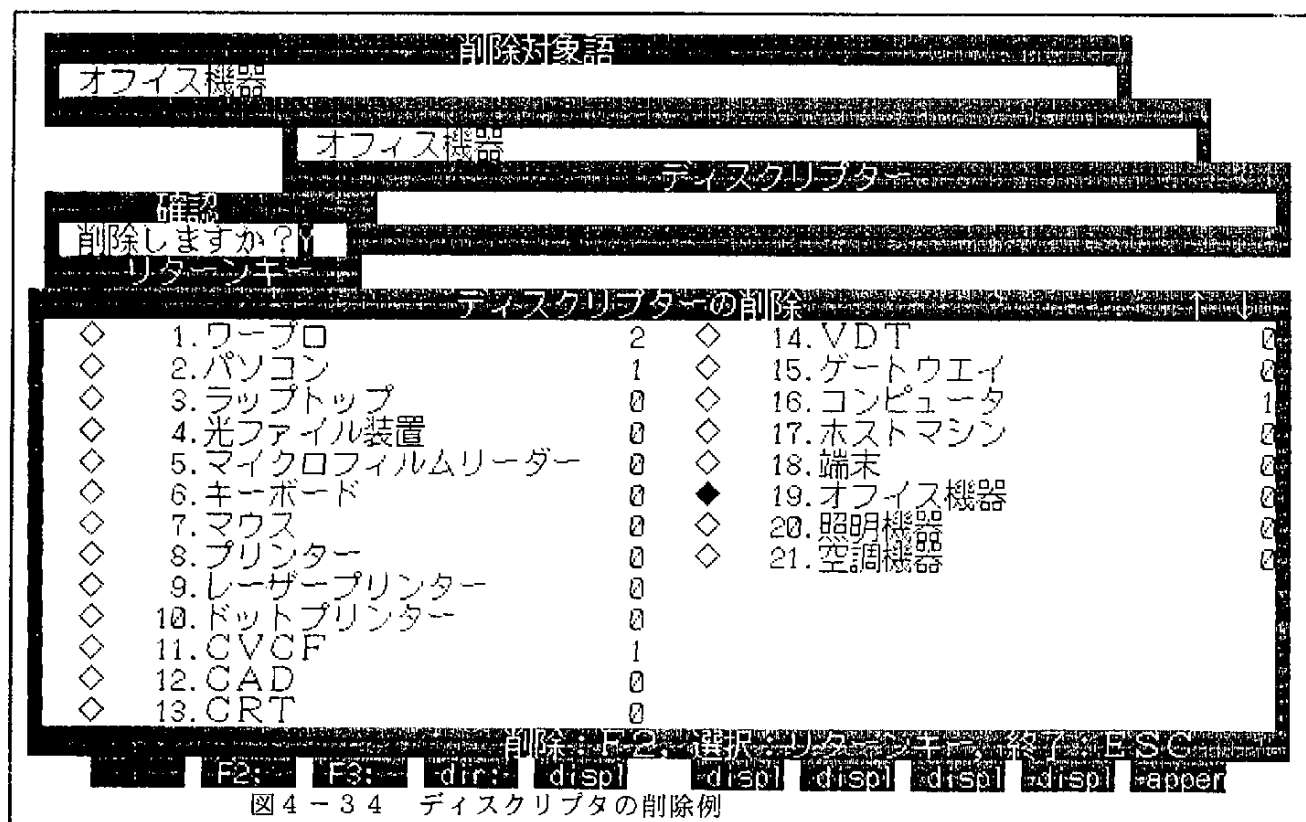


図4-34 ディスクリプターの削除例

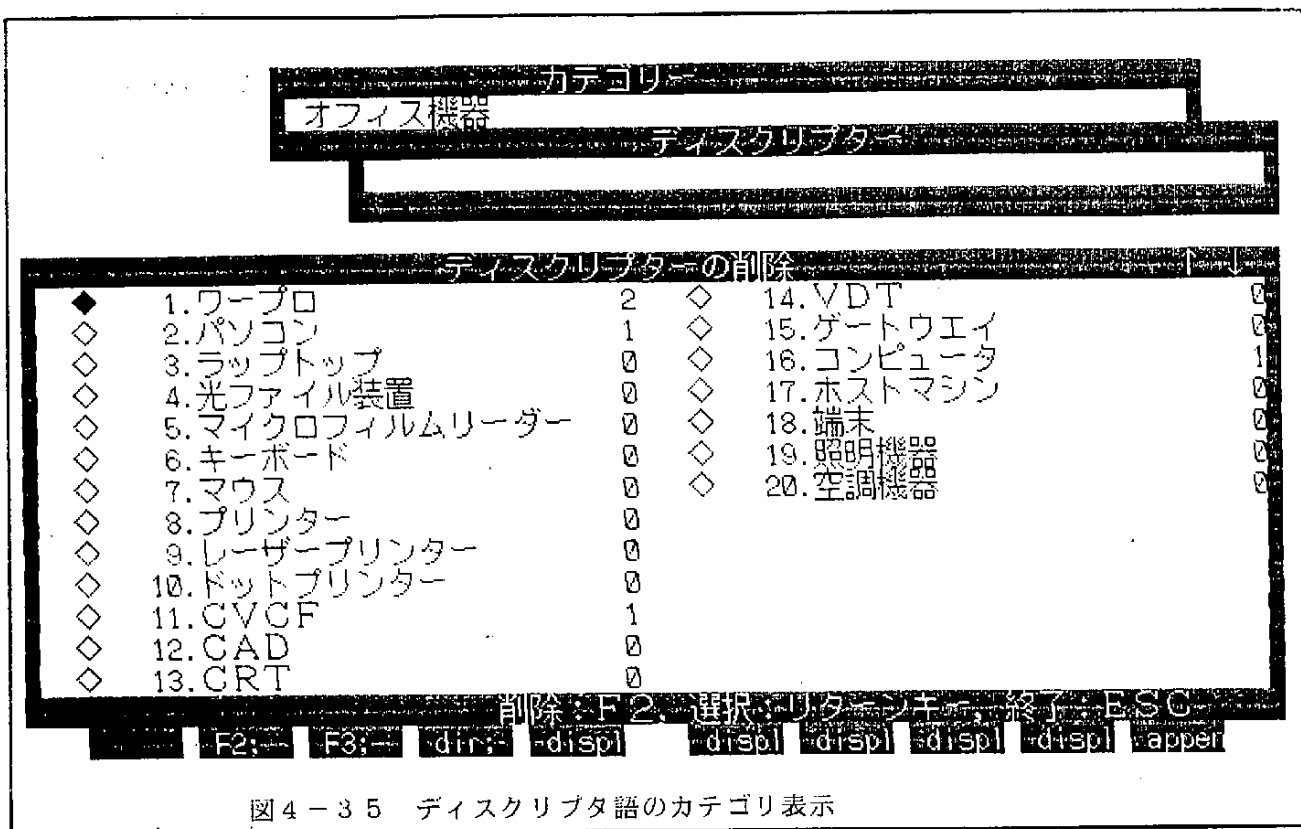


図 4 - 3 5 ディスクリプタ語のカテゴリ表示

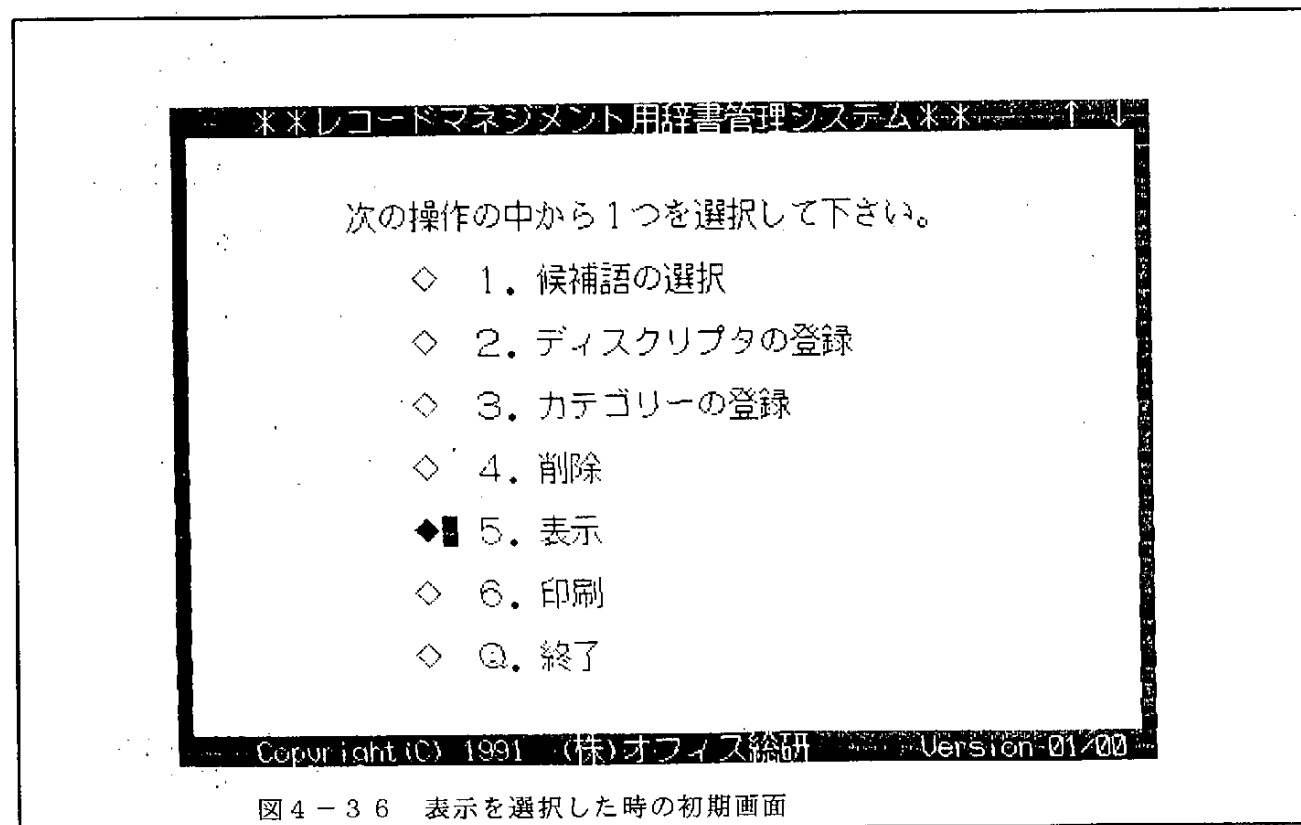


図 4 - 3 6 表示を選択した時の初期画面

ス機器」が削除されている。

なお、カテゴリを削除する際は、その配下にあるディスクリプタや同義語が削除語ファイルに格納される。このファイルは、ディスクリプタの登録処理に利用できるものである。これによりカテゴリの改編（統合、分類等）があったとしても、再度すべてのディスクリプタを入力することなく容易に行うことができる。

4.2.5 表 示

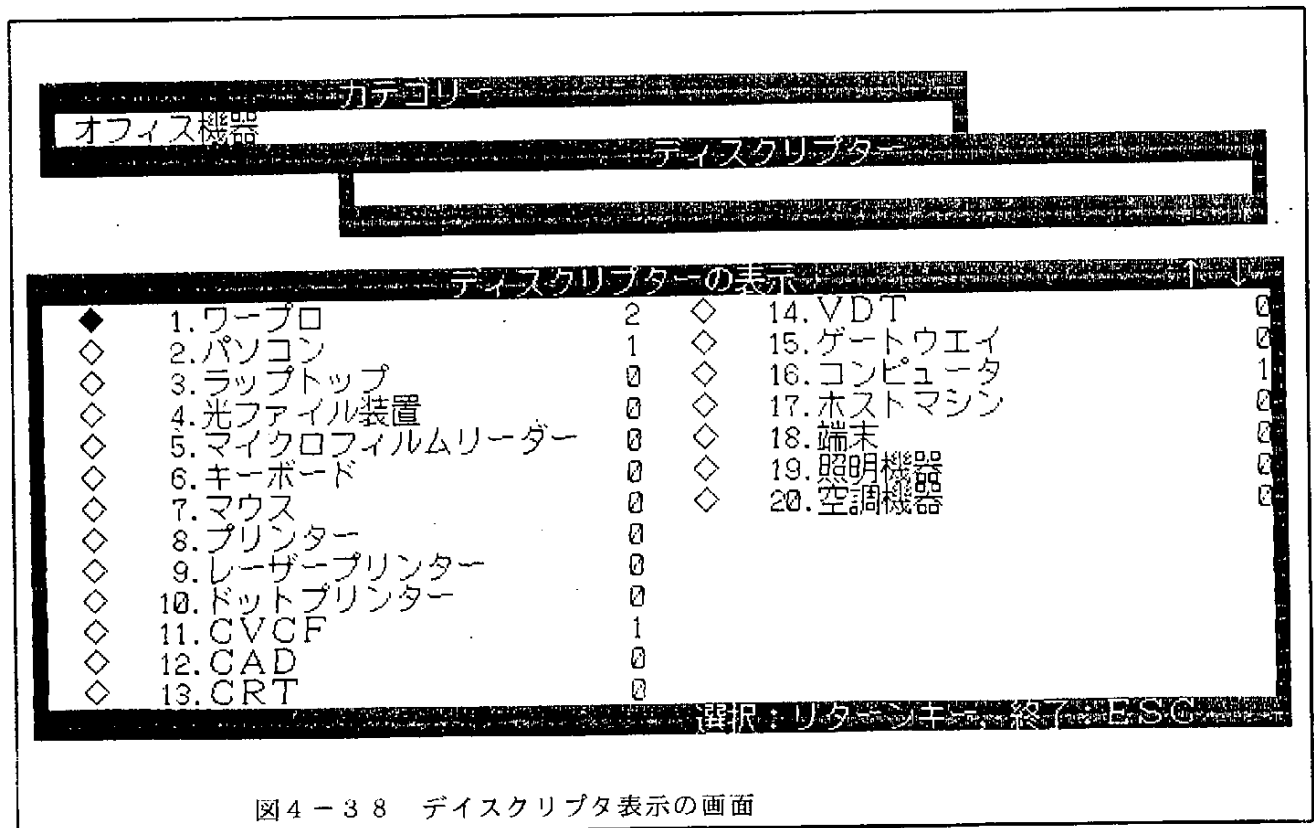
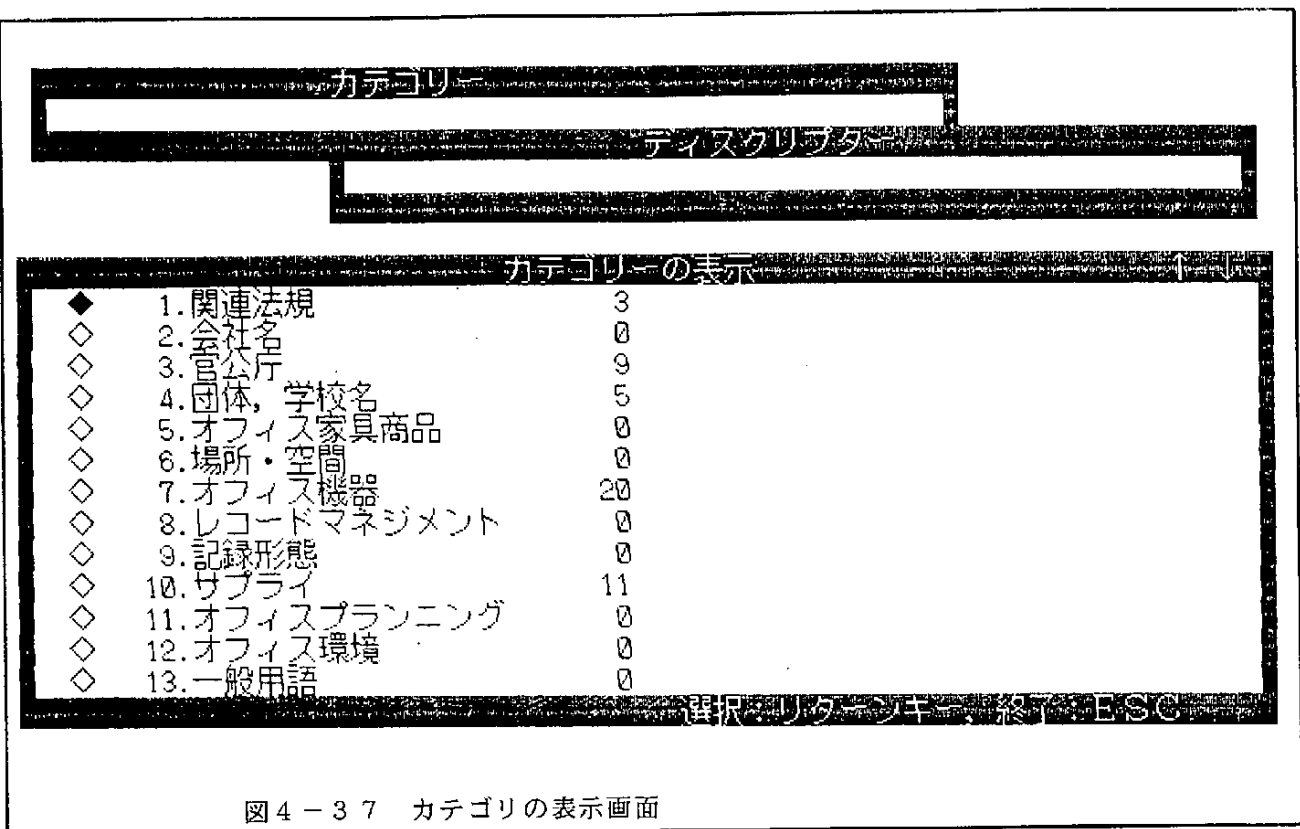
カテゴリ、ディスクリプタ、同義語をディスプレイに表示するだけの機能である。

図4-36の初期画面で5つの表示機能が選択された時に有効となる。

表示画面は、カテゴリウィンドウ、ディスクリプタウィンドウと表示ウィンドウの3つから構成されている。最初にあらわれる画面は、図4-37カテゴリ表示画面であるが、カテゴリは最上位に位置するため、カテゴリ、ディスクリプタウィンドウには何も表示されない。表示ウィンドウにカテゴリが表示されるだけである。ディスクリプタをみたい時は、それが見たいカテゴリを選択することによって図4-38のようなディスクリプタ表示画面となる。この時は、どのカテゴリの配下のディスクリプタであるかを表示するために、カテゴリウィンドウに選択されたカテゴリ名が表示される。さらに、同義語が見たい時は、見たいディスクリプタを選択すると、図4-39の同義語の表示画面となる。その際には、どのカテゴリのどのディスクリプタの同義語であるかがわかるように各ウィンドウカテゴリ名、ディスクリプタ名が表示される。この例は、同義語が存在しない時の例である。

4.2.6 印 刷

シソーラス辞書に登録されている、カテゴリ、ディスクリプタおよび同義語すべての見出しを関係づけた一覧リストを出力する機能である。初期画面で6つの機能を選択すると、図4-40のような確認ウィンドウが表示される。この時点で終了したければ「ESCキー」を、印刷を開始したければ「リターンキー」を押すことによって、プリンタに出力される。プリンタに出力されたカテゴリ、ディスクリプタおよび同義語の一覧リストの例は、表4-1のとおりである。



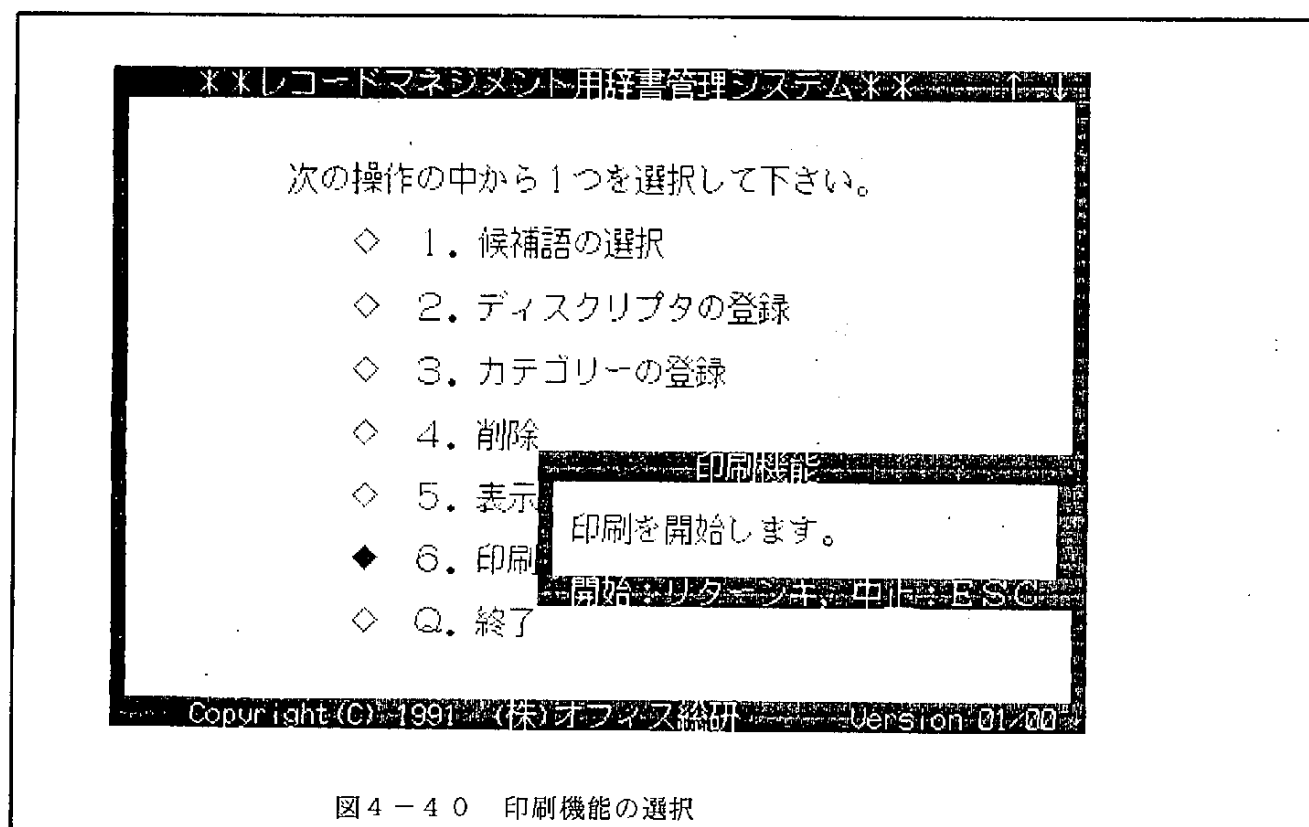
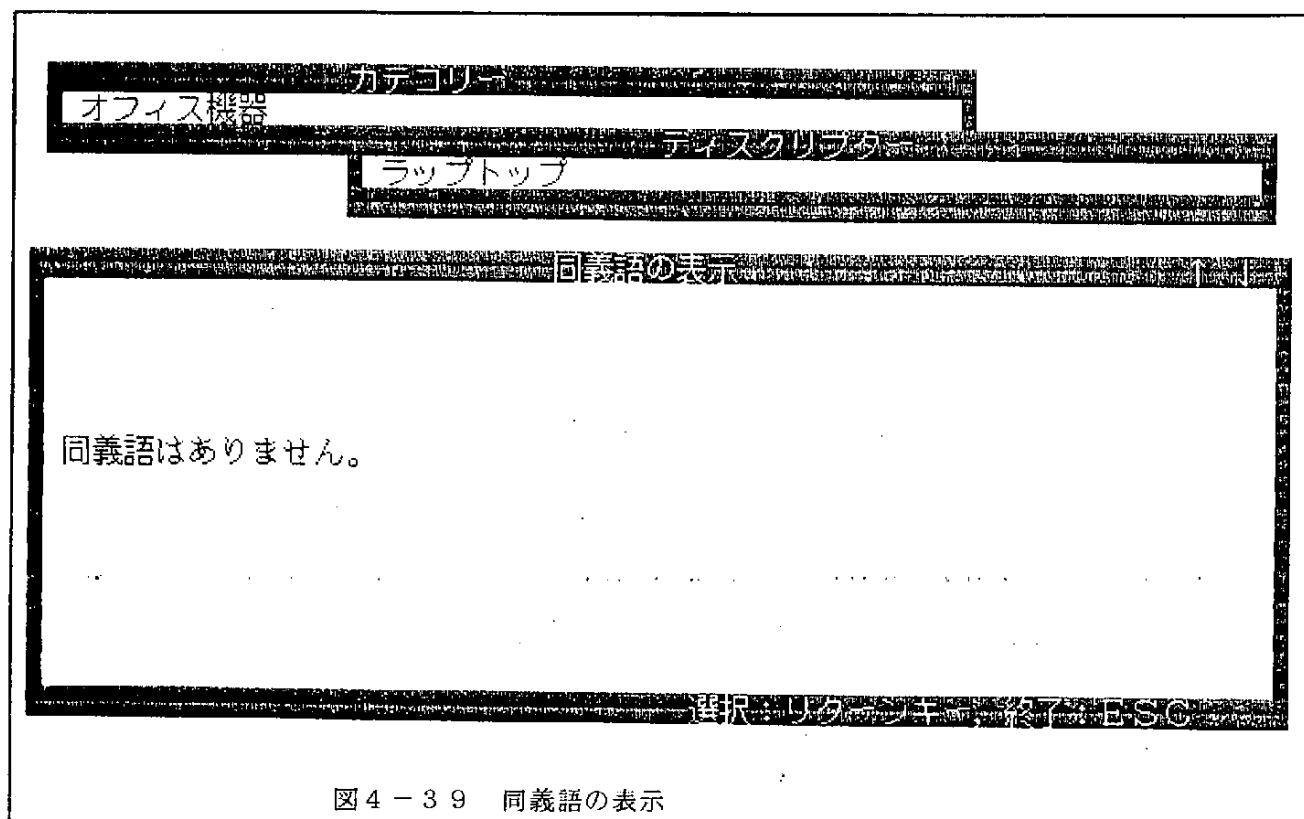


表4-2 シリーラス辞書の一覧リスト

カテゴリー	ディスクリプター	同義語	PAGE. 1
1. 関連法規	1. 建築基準法 2. 消防法 3. 工業標準化法		
2. 会社名	1. 太陽神戸三井銀行 2. オフィス総研 3. O R I 4. オフィス総合研究所 5. 東京レコードマネジメント 6. 関西レコードマネジメント 7. オフィスマネジメントシステム 8. ファイリングクリニック 9. スチールケース 10. オカムラ 11. くろがね 12. I T O 13. 兵庫リコー 14. 京都リコー 15. 岐阜リコー 16. 東北リコー 17. 長崎リコー 18. 愛知リコー 19. 神奈川リコー 20. リコー 21. リコーシステム開発 22. 三愛不動産 23. リコーテレテック	1. T R M 1. K R M 1. 日本スチールケース 1. アイティオー 1. R I S 1. R T T	

4.3. モジュール構成

本システムのプログラムの構成図は次のとおりである。

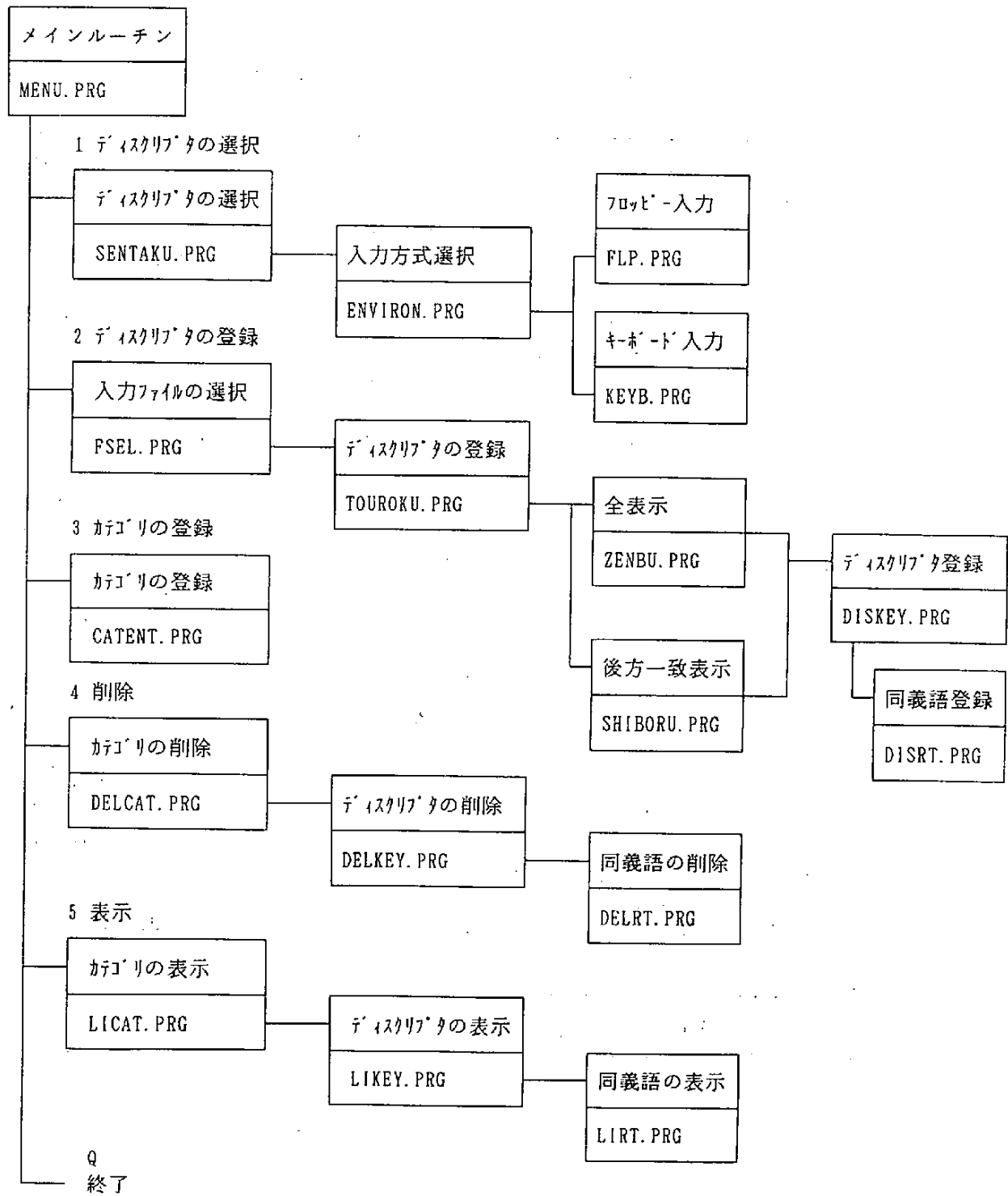


図4-41 レコードマネジメント用辞書管理システムモジュール構成図

4.4. 辞書の構成

本システムで用いている辞書（ファイル）の一覧は次のとおりである。

表4-3 レコードマネジメント用辞書管理システムの一覧

項番	ファイル名	辞書（ファイル）名	インデックス
①	TEXTDB. DBF	候補語辞書	
②	FREQDB. DBF		FTDX. NDX
③	KOUHODB. DBF	ディスクリプタ候補語辞書	
④	SAKUJODB. DBF	削除語辞書	
⑤	CATDB. DBF	カテゴリ辞書	CATDX. NDX HCDX. NDX
⑥	KDB. DBF	ディスクリプタ辞書	KTDX. NDX KITDX. NDX KCDX. NDX HKDX. NDX
⑦	RDB. DBF	同義語辞書	RTDX. NDX RKDX. NDX HRDX. NDX
⑧	WCATDB1. DBF		WCATDX. NDX
⑨	WCATDB2. DBF		
⑩	WKDB. DBF		
⑪	WRDB. DBF		

5. 実験・評価

5.1. 実験

実験の目的は今回開発したシステムの改善点を整理することであり、サンプル文書を使用して運用テストを行った。テストの概要はつぎの通りである。

サンプル文書から単語を抽出した。

その抽出したものを辞書管理システムに入力し、システム上で一定の頻度で（しきい値）出現する新単語について用語およびカテゴリの更新を行いシステムの機能の確認を行った。

5.2. 結果及び考察

開発した辞書管理システムの機能を確認した結果は以下のとおりであり、辞書更新から
の人手作業を軽減するための支援システム機能は達成できたといえる。

- ①一語一語の「キーボード入力」と「FD一括入力」の両機能とも達成できた。
- ②候補語の抽出に出現回数「しきい値」により、登録対象語が自動選択表示されることが確認できた。また「しきい値」を画面の指示に従い、キーボードから変更できることも確認できた。
- ③システムの検索機能により”候補語”と”既に辞書に登録されている用語”とが後方一致した場合に、検索された用語を含むカテゴリを今登録しようとしている用語のカテゴリとして優先的に表示する機能を確認できた。例えば、カテゴリが「会社名」で、その下に「神奈川リコー」という用語が登録されており、新規に候補語として「京都リコー」を入力した。その結果「リコー」で「神奈川リコー」と「京都リコー」が後方一致し、この時、「神奈川リコー」を含むカテゴリである「会社名」を表示した。
- ④同一カテゴリ内の用語の追加はカテゴリの範囲内の任意の画面表示場所（カテゴリ別用語画面に表示されるリスト）に挿入する機能を確認できた。

6. 結 論

レコードマネジメントのための自然語検索システムに用いる辞書更新からの人手作業を軽減するための支援システム機能を開発した。また、その対象となる機能は次の三点である。

- ①候補語の入力を一語一語入力する方法と、フロッピーディスクからの大量一括入力する方法の両方の入力を可能にする。
- ②候補語の出現回数により、辞書登録対象語を自動選択表示する。
- ③辞書更新時、登録対象語がどのカテゴリの配下に属するかシステムの的に判断し、可能性のあるカテゴリを画面に優先表示する。

システムでできることはシステム化（①，②，③）を行った。特に、③のシステム処理（カテゴリ内の既登録語と新規登録対象語との照合）においては、後方一致機能により複合語の照合を達成した。

形態一致機能と意味的一致機能は必要と考えるが、この課題は今後の世の中の科学技術の進歩を待たなければならないものとする。

今後予想される効果として、本システム機能と自然語自動抽出機能を持つ検索システムと組み合わせることにより一般ユーザーでもパーパーファイリングおよび電子ファイリングにおける自然語検索システムの構築と維持が円滑になり、『レコードマネジメントの確立』に大きく貢献できるものと思われる。

調 査 文 献 一 覧

文献No	タ イ ト ル	著 者	掲載本(雑誌) 巻-号	年
1	A CONCEPTUAL FRAMEWORK FOR AUTOMATIC AND DYNAMIC THESAURUS UPDATING IN INFORMATION RETRIEVAL SYSTEMS	M.F.Brundet	COLING 80	1980
2	The Application of a minicomputer to Thesaurus Construction	E.J.Kazalauskas T.D.Holt	Journal of the American for Society Information Science 31- 5	1980
3	Integration of Thesauri in the Social Sciences	Jean Aitchison	International Classification 8- 2	1981
4	I D I O M方式の辞書構成法	山崎正樹, 小橋史彦, 萩野輝雄	電気通信研究所 研究実用化報告 30-11	1981
5	J I C S T 科学技術用語シソーラス 今後の問題	菅宮和夫	情報管理 23-11	1981
6	J I C S T 科学技術用語シソーラス タームチャート—構成、機能、メンテ ナンス—	島海翔	情報管理 25- 9	1982
7	J I C S T 科学技術用語シソーラス 1987年版の作成	斎藤和夫, 坂上安彦, 戸塚隆之, 岡野弘, 杉山邦彦, 五味潤豆	情報管理 9-11	1987
8	Merging thesauri: Principles and Evaluation	Hafedh Mili Roy Rada	I.E.E.E. transaction on pattern Analysis and Machine Inteligence 10-2	1988
9	シソーラスの比較評価 — 「概念体系の 提示」の性能を中心に	岸田和明, 武者野小路恵子, 樋口恵子, 堀垣幾世子, 上田修一, 松山典子	情報の科学と技術 38-10	1988
10	DYNAMIC THESAURAL SYSTEMS: A BIBLIOMETRIC STUDY OF TERMINOLOGICAL AND CONCEPTUAL CHANGE IN SOCIOLOGY AND ECONOMICS WITH APPLICATION TO THE DESIGN OF DYNAMIC THESAURAL SYSTEMS	Lorna K.Rees- Potter	Information processing & Managemant 25- 6	1989
11	J I C S T 科学技術用語シソーラス (特集: 情報検索におけるシソーラスと用語管理)	岡野弘行	情報の科学と技術1 39-12	1989
12	新聞記事情報とN K - M E D I A シソー ラス (特集: 情報検索におけるシソーラスと用語管理)	村田雅博	情報の科学と技術 39-12	1989
13	技術データベースの開発	荒井重信, 中村正則	情報科学技術研究集会発表 論文集 24	1988
14	自然語—統制語変換辞書作成システム	岡野弘行, 坂上安彦, 白山邦彦, 五味潤豆, 戸塚隆之, 佐藤誠, 清水英昭	情報科学技術研究集会発表 論文集 25	1989

<参考文献>

以下に示したものは、調査文献としてその概要を本論文に掲載したもののほかに参考とした文献である。

1. 稲垣博人 文書内容検索システムにおける現状と課題, 情報管理 Vol.33 No2,
pp.124, (1990)

2. (財)データベース振興センター 災害情報シソーラスの構築, データベース構築促進
及び技術開発に関する報告書(01-開-03), p31, (1990)

———— 禁 無 断 転 載 ————

平成3年 3 月発行

発 行 財団法人 データベース振興センター
東京都港区浜松町二丁目4番1号
世界貿易センタービル7階
TEL 03-3459-8581

委託先 株式会社 オフィス総研 総合研究所
東京都中央区日本橋人形町3-10-1
かしきち人形町ビル5階
TEL 03-3664-8841

印刷所 株式会社 青松社
東京都港区南青山2-3-4
TEL 03-3479-7111

