

60—S 002

自然言語のコンピュータ解析用辞書
に関する調査研究報告書

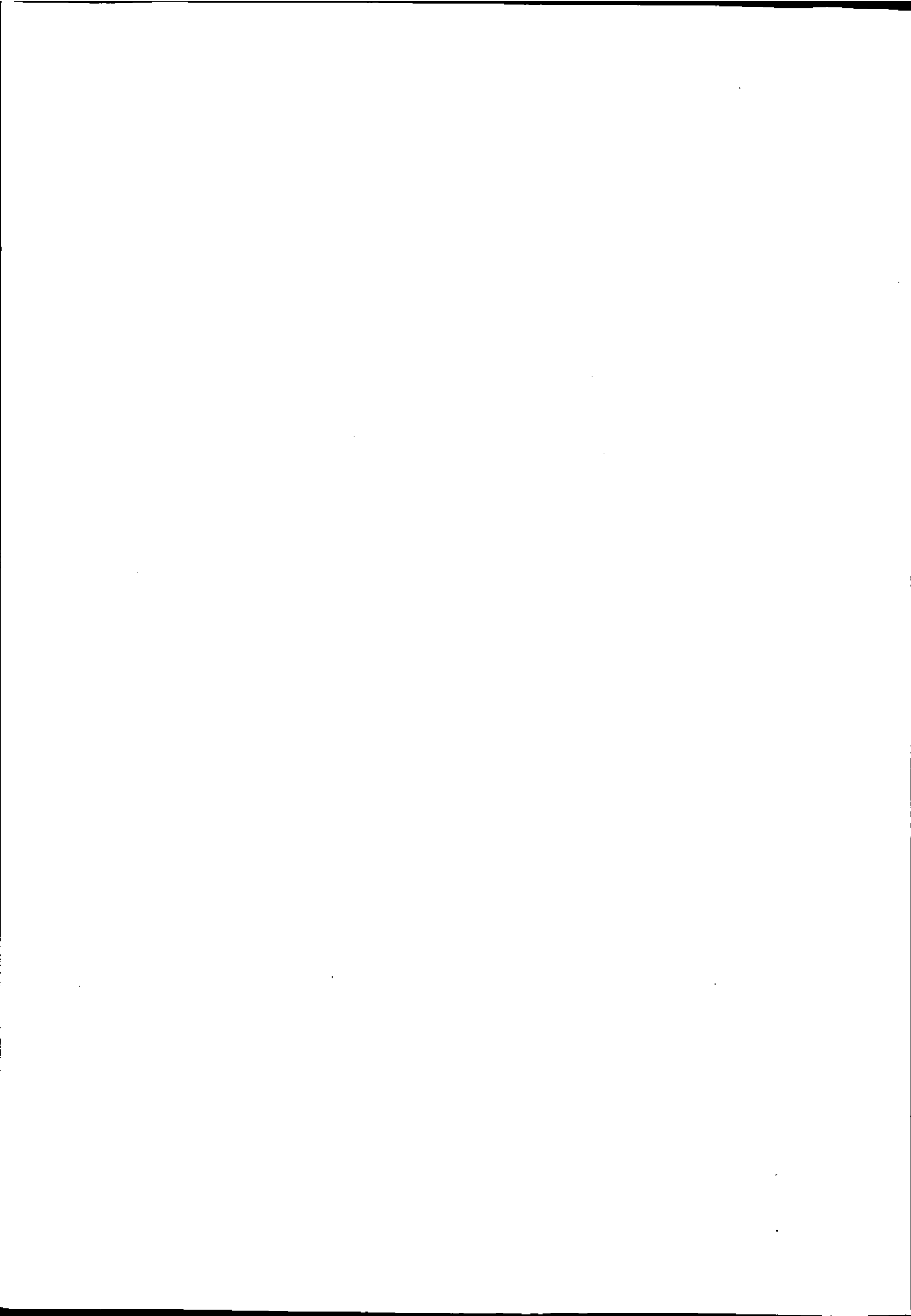
昭和61年 3 月



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60年度に実施した「自然言語の
コンピュータ解析用辞書に関する調査研究」
の成果をとりまとめたものであります。

請求 番号		HPDEC 60-S002		登録 番号	
著者名					
□ 名 自然言語のエンコーダ解析用辞書に 関する言語研究報告書					
所属	帯出者氏名	貸出日	返却 予定日	返却日	



は　じ　め　に

我が国の産業を取り巻く国際経済環境は複雑化してきており、政府、産業界では海外の産業経済動向、技術動向等について迅速かつ正確に把握する必要性が従来にも増して高まっている。

一方、海外の新聞情報、文献情報等は国際ネットワークを通じて即時に入手可能な状況になりつつある。

このような海外の自然言語による文章情報の蓄積、検索利用等の需要が増大する中で、自然言語を対象としたコンピュータによる解析を正確に行うニーズが高まっている。

その際に必要になるのが辞書であり、解析に用いる辞書の良否が解析の精度を決定するとさえ言われている。

現在、自然言語解析の応用技術として、情報の蓄積面では自動インデクシング、自動抄録等に関して、また、情報の利用面では機械翻訳、内容検索、自然言語による質問応答等に関するシステムの研究開発が諸機関において進められているものの、解析に用いる辞書は諸機関における解析手法、解析レベル等の違いにより各々異なっており、システム間の相互利用は不可能な状況にある。

このような背景から、当財団では、辞書のシステム間相互利用を可能とすることを目的に辞書体系を整備し、効率的な辞書作成、維持体制等の確立を図るべく、昭和60年度より3年計画で、自然言語のコンピュータ解析用辞書に関する調査研究を実施することとした。

初年度に当る今年度は、海外も含めた辞書事例調査、それをふまえた辞書の共通要素、各種体系に関する問題点の抽出等を行った。本報告書は、その成果をまとめたものである。

本報告書が、関心のある方々に広く利用され、今後の自然言語処理および辞書研究の向上の一助として寄与できれば幸いである。

最後に、本調査研究にあたって、ご指導ご協力いただいた調査委員長、石川主査をはじめ、各委員並びに関係各位に感謝の意を表します。

昭和61年3月

「自然言語解析用辞書調査委員会」委員名簿

(順不同, 敬称略)

委員長	淵 一博	財新世代コンピュータ技術開発機構 研究所長
委員	国広 哲弥	東京大学 文学部言語学科 教授
"	田中 穂積	東京工業大学 工学部 情報科学科 助教授
"	矢田 光治	財C S K総合研究所 社長
"	瀬谷 重信	日本電信電話㈱ 関東総支社 新規事業開発室 室長
"	石川 徹也	図書館情報大学 図書館情報学部 助教授
"	荻野 綱男	埼玉大学 教養学部 専任講師
"	横山 晶一	通商産業省工業技術院 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 主任研究官
"	山本恵美子	ソフトウェア研究会 副会長
"	中瀬 純夫	リソース・シェアリング㈱ 開発技術第2部 部長
"	木南公統司	㈱開発計算センター システム第二部 課長
"	長谷川 亨	㈱ソフトウェア・リサーチ・アソシエイツ システム本部 システム第5部 部長
"	竹内 憲	日本タイムシェア㈱ 第1システム事業部 部長付
"	藤森 半子	通商産業大臣官房 情報管理課政策情報システム室 室長
"	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
"	佐内 則夫	㈱日本情報処理開発協会 開発部 部長

「自然言語解析用辞書調査専門委員会」委員名簿

(順不同、敬称略)

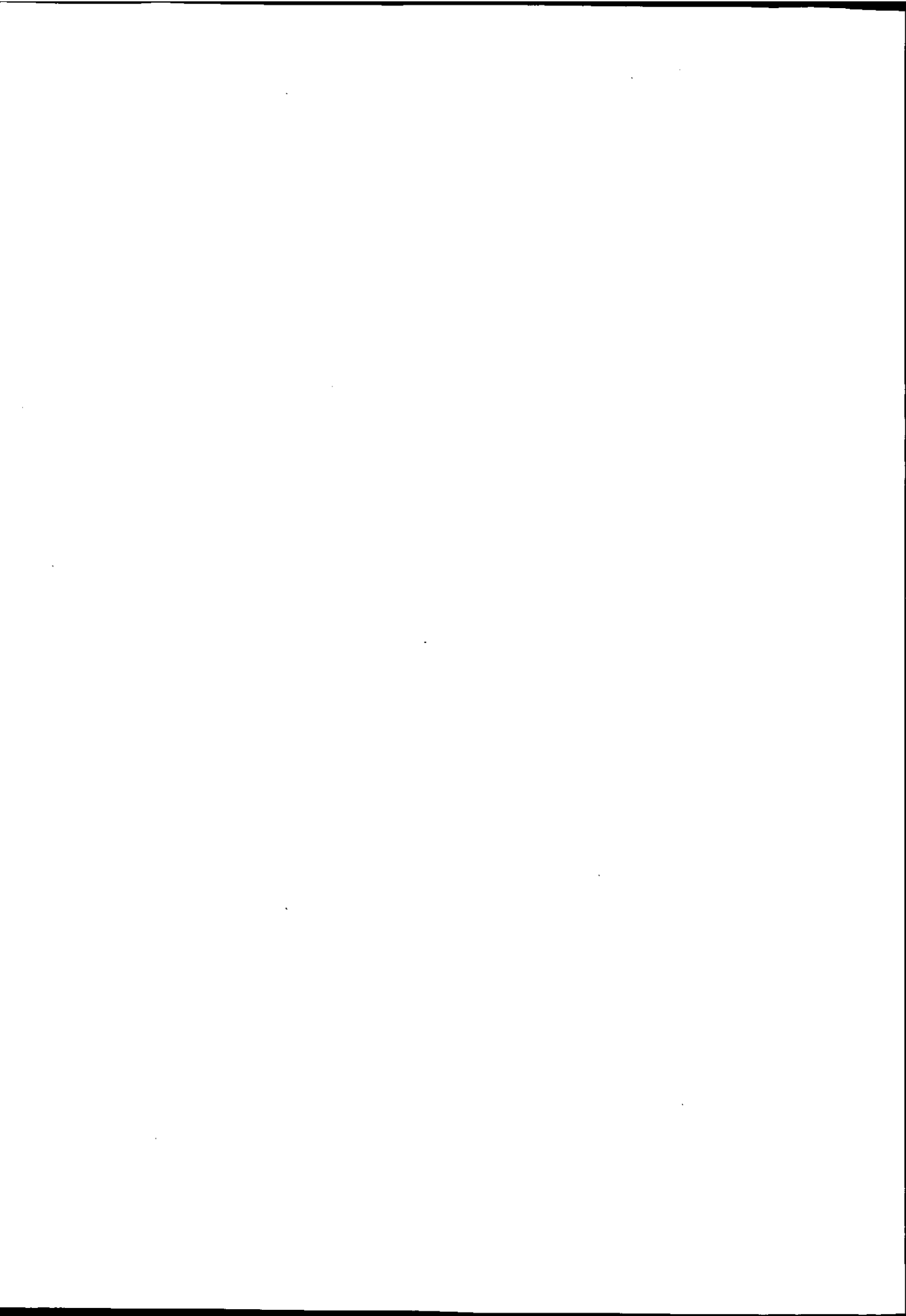
主査	石川 徹也	図書館情報大学 図書館情報学部 助教授
委員	荻野 綱男	埼玉大学 教養学部 専任講師
"	村木新次郎	国立国語研究所 言語体系研究部 第1研究室 室長
"	田中 裕一	㈱新世代コンピュータ技術開発機構 研究所 第二研究室 研究員
"	横山 晶一	通商産業省工業技術院 電子技術総合研究所 パターン情報部推論システム研究室 主任研究官
"	中頼 純夫	リソース・シェアリング㈱ 開発技術第2部 部長
"	木南公統司	㈱開発計算センター システム第二部 課長
"	長谷川 亨	㈱ソフトウェア・リサーチ・アソシエイツ システム本部 システム第5部 部長
"	竹内 憲	日本タイムシェア㈱ 第1システム事業部 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 調査係長
"	鍛冶 勝三	㈱日本情報処理開発協会 開発部 調査役

目 次

1. 事業概要	1
1.1 目的と背景	1
1.2 事業計画と実施経過	1
1.3 調査研究成果の概要	4
1.3.1 本報告書の構成	4
1.3.2 調査研究成果の概要	4
2. 自然言語解析用辞書の必要性	13
2.1 辞書の共用の必要性	13
2.2 辞書共用のレベルと問題点	14
2.3 辞書共用のための環境	17
3. 自然言語解析における辞書の役割	19
3.1 自然言語処理系の構成要素	19
3.2 自然言語解析の各フェーズにおける辞書の利用	20
4. 機械可読辞書の調査	24
4.1 調査項目	24
4.2 海外調査の概要	25
4.3 海外調査の結果	27
4.3.1 西ドイツ情報ドキュメンテーション研究所 (G I D)	27
4.3.2 ボン大学 コミュニケーション・音声研究所 (I K P)	30
4.3.3 ザールランド大学	34
4.3.4 シュトゥットガルト大学	39
4.3.5 西ドイツ言語研究所 (I D S)	42
4.3.6 EC翻訳センター (E C A T)	43
4.3.7 国際用語学情報センター (I N F O T E R M)	48
4.3.8 ビサ大学	53
4.3.9 ウォータルー大学	58
4.3.10 まとめ	64
4.4 国内調査の概要	65

4.5 国内調査の結果	67
4.5.1 アンケート調査結果	67
4.5.1.1 プロジェクトおよび辞書システムの概要	67
4.5.1.2 辞書の概略およびその内容	68
4.5.1.3 辞書の作成およびメンテナンス	69
4.5.1.4 辞書システムの運用	72
4.5.1.5 辞書の共用使用	72
4.5.1.6 参考資料	75
4.5.2 ヒアリング調査結果	77
4.5.2.1 シストラン㈱	77
4.5.2.2 国際電信電話㈱	78
4.5.2.3 株式会社東洋情報システム	78
4.5.2.4 九州大学(吉田 将教授)	79
4.5.2.5 福岡大学(首藤公昭教授)	81
4.5.3 国内調査結果の関連図表	83
5. 自然言語解析用辞書作成の現状	179
5.1 辞書の目的および作成形態	179
5.1.1 汎用の辞書	179
5.1.2 特定のシステムに組み込まれた辞書	181
5.1.3 共用化のポイント	181
5.2 記述内容	181
5.2.1 形態情報	181
5.2.2 音声情報	182
5.2.3 文法情報	182
5.3 意味論的情報および語用論的情報	182
6. 自然言語解析用辞書の記述範囲	184
6.1 自然言語解析用辞書の全体像	184
6.2 共用可能な辞書項目の範囲(英語)	185
6.3 共用可能な辞書項目の範囲(日本語)	191
7. 自然言語解析用辞書の運用	196
7.1 辞書の共用	196
7.2 辞書運用上の考慮点	197
7.3 辞書データ管理システム	198

7.4 辞書の共用化推進のために	199
8. おわりに	200
資料編	202
資料1. 冊子体辞書の歴史と編纂方法	202
資料2. ICOTにおける辞書研究	212
資料3. 機械翻訳システムの評価と使用法	215
資料4. 自然言語解析用辞書アンケート調査票	219



1. 事業概要

1.1 目的と背景

我が国の産業を取り巻く国際経済環境は複雑化してきており、政府、産業界では海外の産業経済動向、技術動向等について迅速かつ正確に把握する必要性が従来にも増して高まっている。

一方、海外の新聞情報、文献情報等は国際ネットワークを通じて即時に入手可能な状況になりつつある。

このような海外の自然言語による文章情報の蓄積、検索利用等の需要が増大する中で、自然言語を対象としたコンピュータによる解析を正確に行うニーズが高まっている。

その際に必要になるのが辞書であり、解析に用いる辞書の良否が解析の精度を決定するとさえ言われている。

現在、自然言語解析の応用技術として、情報の蓄積面では自動インデクシング、自動抄録等に関して、また、情報の利用面では機械翻訳、内容検索、自然言語による質問応答等に関するシステムの研究開発が諸機関において進められているものの、解析に用いる辞書は諸機関における解析手法、解析レベル等の違いにより各々異なっており、システム間の相互利用は不可能な状況にある。

従って、辞書のシステム間相互利用を可能とすることを目的に辞書体系を整備し、効率的な辞書作成、維持体制等の確立を図る必要がある。

1.2 事業計画と実施経過

(1) 事業計画

1.1 で述べた目的と背景から、本事業では、60年度から3カ年計画で、各手法、各機関で相互利用可能なコンピュータ解析用辞書の作成、共同利用等に関して、以下の調査研究を行う。

〔60年度〕

- ・解析用辞書に関する海外も含めた先進事例調査
- ・辞書の相互利用を可能とするための辞書の共通要素を明確化
- ・解析用辞書の全体構成、品詞分類、意味分類等各種体系について検討

〔61年度〕

- ・解析用辞書整備のための基本構想（辞書体系、整備方法、共同利用等）を作成
- ・辞書作成サポートツールの研究・設計

〔62年度〕

- ・辞書作成サポートツールの開発
- ・限定した分野における解析用辞書の整備および実用化のための評価

本事業の初年度にあたる本60年度は、当協会職員および外部有識者からなる「自然言語解析用辞書調査委員会」を設置して、基本方針および実施計画等を検討するとともに、「同専門委員会」において詳細な調査および検討を行うこととした。

なお、海外の先進事例調査にあたっては、調査員を欧・カナダに派遣して実施することとした。

(2) 実施経過

本年度は、辞書事例の調査に力点を置き、委員会および専門委員会における調査検討、国内外の辞書事例調査を行い、それらの調査結果を元にして、辞書の共通要素の明確化、辞書作成・運用などの辞書体系検討のための問題点の抽出を試みた。

なお、海外調査については石川徹也専門委員会主査を調査員として派遣して実施し、国内事例調査については、アンケート、ヒアリングにより調査を実施した。

本年度の調査研究実施経過の概要を表1-1に示す。

表1-1 調査研究実施経過

委員会	回	開催日	議 事
本委員会	1	60. 7. 4	① 委員紹介 ② 事業計画の検討
	2	60.12.19	① 調査研究の実施経過 ② 内外事例調査の実施状況報告 ③ 昭和60年度報告書の骨子
	3	61. 3.10	① 昭和60年度調査研究報告書(案)の検討
専門委員会	1	60. 7.22	① 委員紹介 ② 調査研究計画 ③ 調査研究方針
	2	60. 9. 4	① ヒアリング(東京都立大教授 忍足 欣四郎氏) 【冊子体辞書の歴史と編纂方法】 定義, 歴史, 種類, 語義分析等について, 主として冊子体辞書の作成経験をまじえて解説。 ② 調査事例の選択
	3	60. 9.26	① ヒアリング(ICOT研究所室長 横井 俊夫氏) 【ICOTにおける辞書研究】 ICOTにおける先進的かつ本格的電子化辞書の研究について, 主として推進体制面から解説。 ② 調査項目の検討
			60.10.12 ~ 60.10.27海外調査の実施
			60.10.26 ~ 61. 2.17国内事例調査の実施
	4	60.11.11	① ヒアリング(電力中央研究所 寺野 隆雄氏) 【機械翻訳システムの評価と使用法】 電中研で行った機械翻訳システムの評価実験の紹介とその結果に基づき使用法を解説し, 専門語辞書の重要性に言及。 ② 海外調査概要報告(石川徹也 主査)
	5	60.12. 5	① 国内事例調査進捗報告 ② 昭和60年度報告書構成(素案)の検討
	6	61. 1.20	事例調査結果(表)に基づく検討
	7	61. 2. 3	事例調査結果(まとめ)に基づく検討
	8	61. 2.24	報告書素稿の検討
	9	61. 3. 6	〃
	10	61. 3. 7	〃
	11	61. 3.10	昭和60年度調査研究報告書(案)の検討

1.3 調査研究成果の概要

1.3.1 本報告書の構成

本報告書は、8つの章と資料編から成っている。

第1章は、本事業の概要を述べたものであり、各章の要約も述べられている。

第2章は、自然言語解析用辞書とその共用の必要性について述べられている。(文責 横山晶一)

第3章は、自然言語処理系における辞書の位置づけ、および自然言語解析の各フェーズにおける辞書の利用につき概観している。(文責 中瀬純夫)

第4章は、自然言語解析用辞書を広義にとらえて、機械可読辞書(MRD: Machine Readable Dictionary)の内外の事例について調査した、その概要と調査結果が述べられている。(文責

4.2, 4.3 石川徹也)

第5章は、自然言語解析用辞書を、目的、利用形態、規模、記述項目の種類、記述内容・方法、作成形態等の分類基準により辞書作成について考察したものである。(文責 田中裕一)

第6章は、自然言語解析用辞書の共用可能性の観点から、英語と日本語の辞書項目の記述範囲について考察したものである。(文責 6.1, 6.2 荻野 綱男, 6.3 村木新次郎)

第7章は、自然言語解析用辞書の共用について運用の立場から考察したものである。

(文責 木南 公統司)

第8章は、以上の調査研究成果をふまえて、今後さらに検討を要する主な課題を挙げ、それらの解決のための方策について述べている。

資料編は、専門委員会における講演の概要をまとめた資料1～3と、国内事例調査で用いたアンケート調査票(資料4)より成っている。

なお、本報告書全体を通じて、特にことわりのない限り「辞書」とはコンピュータで読むことの出来る形態の辞書、いわゆる機械可読辞書(MRD: Machine Readable Dictionary)を指す。

電子化辞書も機械可読辞書と同義であるが、固有名詞的用法を除いて本報告書の中では機械可読辞書と呼ぶこととした。

国語辞典などの通常の本の形態の辞書は「冊子体辞書」と称している。

また、機械可読辞書の中でも、本調査研究の対象である「自然言語を対象としたコンピュータによる解析に用いるための辞書」を略して「自然言語コンピュータ解析用辞書」あるいは「自然言語解析用辞書」と称している。

1.3.2 調査研究成果の概要

ここでは、第2章から第7章までの内容を要約して載せ、本調査研究の昭和60年度調査研究成果

の概要を把握しうるようにした。

(1) 自然言語コンピュータ解析用辞書の必要性 (2章)

情報量が著しく増大した現代では、自然言語をコンピュータで解析処理して人間の負担を軽減することが求められている。そのためにも、このような処理のベースとなるべき解析用辞書の充実が望まれる。しかしながらこれには膨大な費用と人手がかかるので、共用化することによってコストの軽減をはかることが望ましい。

2章では、共用の必要性和、その際に生じる種々の問題点について検討した。問題点には、コード体系、ファイル形態、作る立場と使い立場、対象言語や語彙、文法などさまざまなレベルがあり、その内容もそれぞれ大きく異なっている。これら各々の問題点のポイントを指摘した。

最後に、辞書を作るための環境について触れた。共用化をめざした最近の動きについても述べた。また、辞書共用化を進める上で不可欠であるソフトウェアの充実や、ひとまず辞書が完成した後のメンテナンスの問題についても述べた。

(2) 自然言語解析における辞書 (3章)

① 自然言語処理系の構成要素としての辞書に関して、アルゴリズム、規則、辞書の統合体の中で、辞書は各様に規定されていることから、辞書の共用化は、アルゴリズム (さらには規則) の普遍化の観点にもとづきことが必須であることを述べた。

② 自然言語解析の各フェーズにおける辞書に関して、以下のとおり述べた。

(ア) 形態素解析

語の認定および形態素レベルでの各種個別処理のための辞書データについてふれている。

(イ) 統語論的解析

表層構造の context free rule による解析およびそれより深い構造における selective restriction による評価における問題点、主に焦点をあてて述べている。

自然言語解析用辞書のための努力のほとんどは、このレベルとよりアプリケーション依存のレベルとの内容に費やされる。

(ウ) 意味論的解析

辞書における意味記述の形式化は、いまだ実用段階を展望できる状況にはないが、本格的な試行を開始すべき重要な課題である。

(エ) 外部辞書

自然言語処理系の開発運用のための素材、ツールとしての外部辞書データの重要性。

自然言語処理系のヒューマン・インタフェースにおける各種外部辞書利用機能の重要性。

(3) 機械可読辞書の調査 (4章)

① 調査項目

機械可読辞書について以下に示す調査項目を設定して、海外、国内における辞書事例の調査

を実施した。

表 1-2 調査対象辞書に関する調査項目 (表 4-1 と同じ)

- ① 各プロジェクトにおける機械可読辞書の種類, 及びシステムの種類, 及び各システムの機能概要 (機能図)
- ② 機械可読辞書のデータ項目, 及び各データ項目の機能・目的 (システムとの関連)
- ③ 各データ項目のデータフォーマット (各データ項目の記述形式)
- ④ 機械可読辞書のファイルフォーマット, 及び各データ項目間の論理構造
- ⑤ 現時点での単語収録件数, 及び収録語の種類, 及びそれぞれの目的
- ⑥ 各データ項目の内容構成
 - 【例】・見出し語に対する収録語の種類
 - ・品詞分類
 - ・意味マーカ, 及びその体系
 - ・格フレーム, 及びその体系
- ⑦ 機械可読辞書作成上のデータエン트리・ルール (辞書作成マニュアル, 及びサンプル)
- ⑧ 機械可読辞書管理システムの機能概要 (機能図)
- ⑨ 機械可読辞書の作成, 及びメンテナンス体制
 - ・体制・方法
 - ・人数・資質
 - ・経費
- ⑩ 他プロジェクト, 及びシステムとの共用使用にあたり, 辞書作成, 及びメンテナンスを行っていく上での考慮すべき点
- ⑪ システムの利用における利用者側の機械可読辞書の操作可能部分と操作方法
 - ・利用者メンテナンス可能部分, システム側メンテナンス部分
- ⑫ 提供可能性, 及び提供方法
- ⑬ 関連資料, 参考文献

② 海外調査の要約

前述の目的のために 1980 年 10 月 13 日から 10 月 27 日にかけて表 1-3 に示す 9 機関を訪問し調査を行った。

以下に辞書事例の海外調査概略を記す。

A. 西ドイツ情報ドキュメンテーション研究所 (G I D)

- ・西ドイツにおいては、国レベルの、また自然言語処理用の統一的な機械可読辞書（以下、MRD: Machine Readable Dictionary とする）の開発プロジェクトは現在のところないとのこと。西ドイツにおいては、主に、自動翻訳システムの研究・開発プロジェクトに必要とするMRDがプロジェクト単位で作成されている。

・G I Dの概観

G I D（英語名: Society for Information and Documentation）自身は西ドイツ政府の情報政策を具体的にプランニングする機関であり、1977年に設立された政府機関。

主な活動内容は以下のとおり。

- 1) 情報システムに関する研究・開発
- 2) 技術サービス
- 3) 専門知識の提供および専門家の派遣
- 4) 情報科学および情報活動に関する情報サービス
- 5) 情報およびドキュメンテーション活動機関およびG I Dのメンバー機関へのコンサルティング・サービス
- 6) 州および連邦政府の研究プロジェクトの管理
- 7) 国内および国外機関との調整窓口

B. G I D以外の機関におけるMRD構築プロジェクト

今回訪問調査を行った9機関について、G I Dを除き、MRD構築プロジェクトの目的を、以下の2つに類別できるのではないかと認識した。

1) 自動翻訳用MRD

ザールランド大学、ジュットゥットガルト大学、EC翻訳センター

2) 広い意味での自然言語研究支援ツール

ボン大学、西ドイツ言語研究所、INFOTERM、ピサ大学、ウォータルー大学

自動翻訳用のMRD化の研究・開発は、目的語の同定について、例えば日英翻訳のようにまったく語形の異なりがないわけではないだけに、語根の対関係の明確化と、語尾を中心に研究がなされている印象を受けた。それに対して、特に、ピサ大学、およびウォータルー大学のプロジェクトでは、MRDから、いかに知識データベース化を図るか、という研究が開始されたところであるように見受けられた。

C. 海外調査を終えての感想

- 1) 自動翻訳用MRDについては、国レベルのMuプロジェクト、あるいは企業等のプロジェクトで研究が推進されている。

表 1-3 自然言語解析用辞書 海外調査 訪問機関一覧

No.	機 関 名	訪 問 相 手	国
1.	西ドイツ情報ドキュメンテーション研究所 GID (Gesellschaft für Information und Dokumen- tation MBH)	Dieter Mönch (国際部 部長)	西ドイツ
2.	ボン大学 コミュニケーション・音声研究所 IKP (Institute für Kommuni- cation sforschung und Phonetik der Universität Bonn)	Winfried Lenders (IKP 所長, 言語学部教授) Jan Brustkern (IKP 研究員, Lenders 教授の助手)	西ドイツ
3.	ザールランド大学 Universität des Saarlandes	Harald H. Zimmermann (情報科学科 教授) Krishna Sharma (情報科学科 教授) Edith Kroupa (情報科学科 博士課程 女子学生)	西ドイツ
4.	シュトゥットガルト大学 Universität Stuttgart, Institute für Informatik	羽中田 賢治 教授 (情報研究所)	西ドイツ
5.	西ドイツ言語研究所 IDS (The Institute für Deutsche Sprache)	Wolfgang Teubert (GID 副所長) Klaus Wothke (研究員) 在 岡 進 (東京外国語大学 ドイツ語学科 助教授)	西ドイツ
6.	EC 翻訳センター ECAT (European Center for Automatic Translation)	Miehel Lebrun (ECAT マネジャー) Cay - Holger Stoll (マネージング ディレクター)	ルクセン ブルグ
7.	国際用語学情報センター INFOTERM (International Information Center for Terminology)	Christian Galinski (副所長)	オースト リア
8.	ピサ大学 University of Pisa, Department of Linguistic	Nicoletta Galzolari (言語学科 助教授) Antonio Zampolli (言語学科 教授)	イタリア
9.	ウォータールー大学 Center for the New Oxford English Dictionary	Gayle Johannesen (事務長 女性) J. Stubbs (Univ. Oxford Press の OED 編集部からの出向) F. W. Tompa (計算機学科 助教授)	カナダ

2) 西ドイツ言語研究所のような国語研究のためには、国立国語研究所が研究を推進している。

3) INFOTERM のような Terminology については、JIS 用語あるいは、シリーズの構築の中で推進されていると考える。

以上のように考えると、残るテーマとして次のようなことがあると考えた。

- 1) EURODICAUTOMのように、1つのプロジェクトでは人的にも、資金的にも、時間的にも成し得ないMRDを国のレベルで開発していくことが大切である。
- 2) ビサ大学、およびウォーター大学構想のように、冊子体レベルのMRDから、知識データベース化を図る研究をする必要がある。

特に、前者の1)のレベルについては、国レベルのプロジェクトとして構築を推進することが大切であると思えた。後者の2)のレベルについては、一つのプロジェクトで研究をスタートさせる必要があると思えた。

③ 国内調査の要約

国内調査はアンケート調査とヒアリング調査がおこなわれた(4.4)。アンケート調査は、昭和60年11月26日より昭和61年2月17日まで、24機関26辞書について実施され、19の辞書についてアンケートの回答をよせていただいた(表1-4)。アンケートの内容には、先端的色彩が強く研究開発上の機密に属する事項が多く含まれており、回答に際して、企業戦略上、研究開発上、公表可能な範囲で協力をお願いした。この結果は、アンケート調査の結果(4.5.1)にまとめられている。また、アンケート調査の結果を補うべくヒアリング調査がおこなわれた。ヒアリング調査の訪問先は、シストラン(株)、国際電信電話(株)、(株)東洋情報システム、九州大学吉田将教授、福岡大学首藤公昭教授である。この結果については、ヒアリング調査の結果(4.5.2)にまとめられている。

表1-4 アンケートの回収状況一覧

分類	機械可読辞書	機械可読 辞書の略称	調 査 対 象 機 関	アンケート 回収状況
一般辞書	(1)新明解国語辞典(三省堂)	新明解	工業技術院電子技術総合研究所	回収
	(2)学研国語辞典(学研)	学研	(株)日本電子工業振興協会	回収できず
	(3)ニューコンサイス英和辞典	コンサイス	京都大学	回収
	(4)Longman	LDOCE	京都大学	回収
専門辞書	(5)I P電子辞典シリーズ	インタープレス	(株)インタープレス社	回収
	(6)I S O用語ファイル	I S O	(株)日本規格協会	回収できず
	(7)学術用語ファイル	学術用語	文部省	回収できず
	(8)電字林	電字林	(株)東洋情報システム	回収

表1-4 アンケートの回収状況一覧(つづき)

分類	機械可読辞書	機械可読 辞書の略称	調査対象機関	アンケート 回収状況
目的辞書	(9)Mu 機械翻訳プロジェクト	Mu	科学技術庁	回収
	(10)ICOT電子辞書システム	ICOT	財新世代コンピュータ技術開発機構(ICOT)	回収
	(11)日本語基本動詞辞書	IPA	情報処理振興事業協会(IPA)	回収
	(12)JICSTシソーラス	JICST	日本科学技術情報センター	回収できず
	(13)情報検索用辞書	JICST	日本科学技術情報センター	回収できず
	(14)JAPIOシソーラス	JAPIO	財特許情報機構(JAPIO)	回収
(商用/ 実験) 機 械翻訳シ ステムの 辞書	(15)PIVOT/VENUS	日電	日本電気(株)	回収
	(16)ATLAS-I (英日) ATLAS-II (日英)	富士通	富士通(株)	回収
	(17)TAURAS (英日)	東芝	(株)東芝総合研究所	回収
	(18)メディアムバック・マイ クロバック	ブラビス	(株)ブラビス・インターナショナル	回収
	(19)機械翻訳システム	シャープ	シャープ	回収できず
	(20)ATHENE/E ATHENE/J	日立	(株)日立製作所システム開発研究所	回収
	(21)日英機械翻訳システム	三菱	三菱電機(株)	回収できず
	(22)機械翻訳システムSHALT	日本IBM	日本IBM(株)	回収
	(23)SYSTRAN	シストラン	シストラン(株)	回収
	(24)VALANTINE, KATE	KDD	国際電信電話(株)	回収
	(25)LUTE	NTT	日本電信電話(株)	回収
	(26)機械翻訳システム	沖電気	沖電気工業(株)	回収

(4) 自然言語解析用辞書作成の現状(5章)

共用性という観点から見た自然言語解析用辞書の作成に関わる問題点を扱う。

辞書が依拠する文法の理論や解析のアルゴリズムは共用性に対する障害になる。辞書の記述においては特定の文法理論に拠らざるを得ない部分が相当あるが、そこでは十分一般性のある理論

を選択しなければならない。またデータ構造、データの表現も特定のアルゴリズムによらないものを考慮する必要がある。

作成形態として、冊子体の辞書を機械可読の形式に変換する試みが多くなされているが、人間が読むことを想定して記述された情報を機械が利用することは難しい。むしろ、執筆等々の開発体制は冊子体辞書と同じにして、最初から機械可読用の仕様、内容に基づいて開発するほうが良い。

意味論的、語用論的情報については、現在各所で実験的に研究が進められている状態であり、大規模な汎用機械可読辞書に取り入れることには問題があろうが、これを扱おうとする試みもある。

(5) 自然言語解析用辞書の記述範囲（6章）

① 自然言語解析用辞書の全体像

応用システムとして英日機械翻訳を考え、英語辞書・日本語辞書両方についてあるべき姿を考える。自然言語解析用辞書は、固有名詞や複合語を含めて規模（収録範囲）が数十万語レベルになり、次のような辞書項目を含む。

ID情報：見出し・一連番号など

形態情報：表記・語形変化（動詞の活用など）

統語情報：品詞・共起するもの・連語（慣用句・イディオム）

意味情報：訳語・関連語（上位語・類義語・反対語、将来の課題として全体部分関係など）

② 共用可能な辞書項目の範囲（英語）

応用システムごとのアルゴリズムの違い、文法の不完全性（言語の記述の不完全性）、文法との関連といった問題があるので、共用辞書という考え方はあまり有効でない。一応「事実に見して明示的に記述する」という方針で、共用辞書について考える。

1) 品詞：下位分類はしない。その理由は、下位分類の基準がわからないだけでなく、下位分類をすると「多品詞」語の増大と品詞のカテゴリー数の増大によって全体が見通せなくなるからである。

2) 共起関係：動詞の場合は、文型パターン・主語や目的語にあらわれる名詞・副詞語句・前置詞句。名詞の場合は冠詞（countableかuncountableか）・形容詞・その名詞を修飾する名詞。形容詞の場合は、それを修飾する副詞。

3) 連語：構文情報（全体の品詞と個々の構成要素の品詞・連語性＝結合の強さ・構成要素の置換性など）・意味情報（関連語など）・連語でない通常の意味での用法もあるか…。

4) 意味マーカ：共用辞書には意味マーカを含めるべきではない。というのは、意味マーカの利用方法・目的が多様で、設定の基準がない上に、意味マーカでは意味を明示的に記述するのに粗すぎるからである。

③ 共用可能な辞書項目の範囲（日本語）

辞書の記述内容は、文法理論に依存する部分があり、汎用の辞書を求めるのは容易なことでは

はない。本報告では、基本語彙の記述をめぐって問題となる項目に限ってとりあげた。

見出し語については、単語の認定、合成語の扱い、表記の多様性などの問題がある。見出し語を下位区分する際、語彙的意味の差異だけでなく、文法的な特徴にも充分考慮されなければならない。意味記述の用語の限定、記述方法の統一が達成されるなら機械処理にとって有効である。上位語、下位語、類義語、反義語などの関連語を記述するのは当然であるが、シソーラスと百科事典的な情報は、日常語基本語の場合、分類や記述の範囲が明確となりにくい。文法情報としては、語形変化（広義）、形態論的カテゴリー、支配語による語結合のパターンの記述は必須であろう。これに加えて、文体に関する情報、慣用表現も登録されなければならない。

(6) 自然言語解析用辞書の運用（7章）

言葉を扱うシステムの研究・開発にあたっては、まず何らかの形で自然言語解析用の辞書を用意しなければならない。このようなとき、既存の電子化辞書を元にして目的の辞書作りができれば、開発費用・期間および労力の短縮ができる。このためには関係する専門家あるいは諸機関による辞書共用化のための協調体制が不可欠であり、かつ行政からの支援が望まれる。

一方、ユーザが適用システム上で辞書を運用していく上では、辞書の提供側との協調体制、辞書編集手順標準マニュアルの整備、辞書編集機能、専門家による技術支援などを考慮し適切な運用体制の確立が必要である。また、従来、辞書データベースの管理に汎用のデータベース管理システムが採用されている例が多いが、今後は辞書の管理に適し、アクセス効率や運用上の容易性の高い辞書専用のデータベース管理システムの開発が望まれる。

2 自然言語解析用辞書の必要性

2.1 辞書の共用の必要性

「情報化社会」という言葉が言われ、情報というものが我々にとって欠くべからざるものになって久しいが、最近では逆に情報の過剰、氾濫が問題にされるほど情報の量そのものが増大する傾向にある。このような膨大な情報の中から、必要な時に必要な情報を的確に取り出すことは非常に重要で有益である。そのためには、データベースの構築なども重要であるが、データの多くを占める自然言語をコンピュータ上で扱えることが望ましい。また、データの管理、運営なども自然言語を基本入出力とするシステム上で行うことができれば、ユーザになじみやすいシステムになると考えられる。

このような自然言語処理システムの重要性は早くから指摘され、様々なシステムが研究、開発されてきた。現在話題となっている機械翻訳、自動抄録、データベース検索などのシステムも、広い意味での自然言語処理システムと見なすことができる。

自然言語処理そのものの内部を見た場合、構文解析、意味解析、文脈解析など、自然言語そのものを扱うためのアルゴリズムを研究することも重要であるが、システムの実用化ということに着目した場合には、自然言語の解析（上に挙げたような様々なレベルでの）を支援するための辞書が重要になってくる。これらの辞書は、研究レベルから既にシステムの性能に影響を与えるだけでなく、実用レベルでは、拡張、変更が容易に行えるようなものが望まれる。また、辞書そのものの開発にはかなりのコストがかかる。これまでの多くの自然言語処理システムにおいては、ほとんどの場合、これらの辞書はシステムに依存する形で作られてきた。従ってシステムの開発が終了した時点で、これらの辞書も他に用いられることなく放棄される例がしばしば見受けられる。

一方、自然言語処理システム用に開発されている辞書とは別に、市販されている冊子体辞書を電子化する例も幾つか見られる。このような辞書は、いわゆる「電子化辞書」として、冊子体辞書を使用するのと全く同じ操作をコンピュータの端末から行って普通の意味での辞書として使用できることはもちろんであるが、自然言語処理システムそのものへの寄与もはかりしれないものがある。まず見出し語そのものを自然言語解析用辞書のエントリとして使用することができるし、見出し語の数も豊富である。さらに一国語の辞書（国語辞典、英英辞典など）の場合は、語釈の部分を解析することによって、意味解析などに役立てることができるし、二国間辞書では、機械翻訳の様々なレベルでの利用が考えられる。また、その他にも個々の辞書を見ると、もっと細かいレベルでの利用も可能である。たとえば国語辞典ではカナ漢字変換や漢字カナ変換（いわゆる読み方）に使うこともできるし、アクセントを付した辞書ではそれを用いて音声出力の改良を行うことも考えられる。英和や英英辞典

の場合には、発音部分を同じような用途に使用できるし、和英辞典の場合には見出し語部分を使うと国語辞典と同様の使い方ができる。これらの辞書を組み合わせるとさらに種々の利用法が生まれてくるかもしれない。ここに述べたような意味での電子化辞書は、著作権などの問題もあって二重、三重に作られることは少ないが、逆に著作権がネックになって公開されないとか、使用の際に様々な制限がつけられているものも多い。

以上のような2つの意味での辞書は、現状ではどちらも種々のメーカ、研究所などによって独自に開発されている。従ってこれらの辞書の相互利用の可能性を考えることは、辞書開発に伴う労力、コストの軽減に役立つものと考えられる。

2.2 辞書共用のレベルと問題点

辞書システムの相互利用という観点から考えると、そこには種々のレベルがあり、また様々な困難な問題がある。以下ではそのようなレベル、問題点を指摘してみたい。

(1) コード体系

最も原始的なレベルではコードの問題がある。アルファベットのレベルでも、現状ではASCII, EBCDIC, BCDが混在している(EBCDICには英小文字の部分にカナコードを入れたものもある)。日本語のレベルでは、JISコード体系が規準となつてそれとの互換性をもつシステムが多くなりつつあるものの、実際のコンピュータ上ではJIS設定以前からの様々なコード体系が存在して、データの相互利用を妨げる一因となっていることは否定できない。またJISコード体系の性質から、そのままではASCII, EBCDICなどと混在できないので、各バイト(JISコードは2バイトで一組になって一つの漢字を表わすが、各バイトのうちで実際に使用されているのは下の7ビットである)の最上位の1ビットを1にしたいわゆる「シフトJIS」コードを使用しているコンピュータも多い。これならASCIIとはコードの重なりを防げるので混在が可能であるが、EBCDICと混在させるためにはコードの変わり目ごとに印を入れる必要がある。この手順もJISに規定されているが、字種変更ごとに3バイト必要なので必ずしもすべてのコンピュータがこの手順に従っている訳ではない。さらに、英語や日本語以外の言語を扱おうとすると問題はもっと厄介になる。ロシア文字、ギリシャ文字についてはJISに含まれているので日本語と同じように取り扱うことができるが、ヨーロッパ系の言語をとっても、フランス語のアクサンやセディーユ、ドイツ語のウムラウト、スペイン語の疑問符や感嘆符の取り扱いなど細かい点では問題も多い。その他の系統の言語になると問題はもっと複雑になる。たとえばハングルは字数だけからみるとかなり多いが、一つ一つの要素に分解するとバリエーションはかなり少なくなる。似たような言語はまだ他にもある。またアラビア語系の言語のように右から左へつづるものは表示のときの工夫も合わせて必要になる。このようないろ

いような言語を混在して表示する場合には各々の言語の変わり目で適当なコードを入れて扱うことになるであろうが、その際にはソフトウェア的な手だてが重要になる。

(2) ファイル形態

コンピュータ内部の問題としては、ファイルの形態の問題がある。辞書データは多くの場合、容量的に見て、ディスクや磁気テープなどの媒体に格納されていることが多いが、現状ではこれらの媒体の内容はコンピュータシステムに依存しており、他システムに移行しようとする、ファイル形態そのものが異なっていて困難であることが多い。ここまでの低レベルの話は、多く小さな問題として片付けられがちであるが、実際に使用する際に意外にコストがかかるものである。広い意味での辞書作成サポートツールでは、このあたりのソフトウェアの充実まで考慮する必要があると思われる。また、最近のコンピュータ間の通信の発達から、オンラインで異なる形態のデータのやりとりが容易に行われるようになれば、ハードウェアやファームウェアのレベルでこれらの問題が解決される可能性もある。

(3) 作る立場と使う立場

辞書を作る立場と使う立場（現在までは同じことが多いが、相互利用するという観点から見た場合には分かれる可能性がある）から見た場合には、またそれぞれの問題点がある。どちらの立場から見ても使い易さとかメンテナンスのしやすさを考慮することはもちろんであるが、しばしば両方で立場の違いが出てくる場合がある。たとえば作る側としてはオリジナルな形を改変されることは恐らくあまり好まないであろうが、逆に使う側としては、自分に合った形に辞書を変形したい場合が多い。また辞書の一部分のみを抽出してそれを適当に組み合わせたり、複数の辞書を自分の都合のいいように結合して使用したいこともある。このようにした時の辞書のバージョンアップ、メンテナンスをどうするか、またファイルの仕分けなどの問題点も多い。

また、使う側から見た場合、どのような形態で使用するかも問題になり、それが作る側の態度に影響を与えることもある。たとえば辞書を機械翻訳に使う場合でも、一度プログラムがスタートしたら途中でユーザの介入を許さずに終わりまで走るやり方（バッチ処理に近い）と、コンピュータとユーザがディスプレイ端末を介するなどして常にインタラクティブに対話しながら処理を進めていくやり方とでは、辞書の使い方が異なるのはもちろん、元の辞書の作りそのものが違っている必要がある。ディスプレイ端末も、大型コンピュータに接続しているか、ワークステーションにつながっているか、後者の場合はワークステーション同士、ワークステーションと大型コンピュータ間でのネットワークなどの通信手段があるかによって辞書の形に影響を与えることはありうる。

(4) 対象言語、語彙

対象とする言語によっても項目の設定、ファイル形態、使用するコードなどが違ってくるのが考えられる。品詞分類を例にとれば、英語の辞書には冠詞、前置詞などの項目はあっても日本

語にはないし、日本語の助詞、連体詞は英語にはない。ドイツ語の分離動詞、格変化は英語や日本語にはないし、ヨーロッパ系言語の男性、女性名詞などは英語では一般的でない。また、そもそも品詞分類自体に種々のものがあり、言語学では常に活発な議論の対象となるところである。たとえば日本語では、形容動詞というカテゴリを認めるかどうかは常に問題にされており、冊子体辞書ではこのカテゴリを採用していないものもある。また、名詞の中でサ行変格活用をする、いわゆるサ変名詞は、英語の動名詞のように動詞と名詞の両方の側面を持っている上に語の数も多いのでその取り扱いには注意を要する。その他の言語になれば、すでに述べたコード体系やフォントの問題ももちろんであるが、そもそも文法について基本的な事柄すら明らかになっていない言語が多いなど、さらにまた種々の問題が生じる。

分野によっても語彙のセットは異なる。もちろんその言語を用いる場合に最小限必要な基本語彙といったものは存在するであろう。しかしながら同じ語（これは基本語彙の場合でも起こりうる）でも分野によって意味の異なる使い方をすることがある。これは電気工学、情報処理など、ある程度分野を絞りこんでもかならず起こることである。分野を絞りこむと、こんどは専門用語の取り扱いをしなければならない。専門用語の場合には、語の登録、収集も労力を要するが、その他にも、専門用語としてはよく出てくる長い複合語（多くは名詞であるが）をどう扱うか、どの程度までそういった複合語を認めるかということも考えなければならない。一つの語で多くの意味、使い方をする多義語の場合には使用頻度を考慮した辞書構成も必要であろう。また最近のワードプロセッサの一部に使われているように、処理の際に使われる頻度によって優先度が変化するような、いわゆる「学習」機能を持たせることも必要になってくるであろう。

(5) 文法

一方、文法的なことを考えた場合、その用途によって当然のことながら辞書作りの形態も異なってくる。上に述べたような、言語の違いも関係するが、文法レベルの違いも大きい。形態素レベル、構文解析レベル、意味解析レベル、文脈レベルと、各段階に応じて、辞書の形態は異なったものにならざるをえない。機械翻訳の場合にはこれに加えて二国間辞書、あるいは中間言語形式をとった場合には中間言語との間の辞書が必要になる。さらに問題を複雑にしているのは、確立した文法形式が定まっていないうえに、各々のレベルでも、その方式によって辞書の形が異なってくるということである。たとえば、格文法で意味素性を記述する場合を考えてみると、同じ格文法といっても意味素性の名前や項目の数が違う。同じ事柄を表わすのであれば名前の差異には目をつぶるとしても、同じ語が全く違うカテゴリに分類されたりしては混乱を招くであろう。逆にカテゴリ名が同じでも文法理論によって異なるニュアンスで用いられているものがあり、注意を要する。

2.3 辞書共用のための環境

このように辞書作成の上ではさまざまなレベルが必要である。従って一つの辞書で全体をカバーせずに辞書に階層性を持たせることが自ずから要求される。上に述べた基本語彙と専門用語の階層化も困難ではあるが行う必要があるであろう。その際はたとえば基本語彙とある程度までの専門用語を作成者が提供し、残りの専門用語はユーザが付け加えるといったことが考えられる。ユーザからのフィードバックも行えればなおよいであろうが、その場合には上記のようなメンテナンスの問題がでてくる。

しかしながら辞書に階層性を持たせたり、さまざまなカテゴリをただ無批判に取り入れただけでは、余りにも辞書が細分化され過ぎるのではないかという懸念が生じる。そもそも辞書を細分化すべきか、細分化するとしたらどの程度まで行ったらよいかということについては充分な検討が必要である。また大きな知識ベースを中心にすえて、それからデータを取り出す場合に初めてソフトウェア的な操作を行って用途に合った辞書を作成することも考えられる。このようにした場合には、コンピュータ解析用辞書という呼び名を広く捕えて、もとの辞書周りのソフトウェアまで含めた総合的なシステムを構築することまで考える必要がある。現在電子辞書の作成の際どのような方法論で辞書を作っていたらよいかということについてはいまだに定まった理論というものはない。これらの辞書作成方法論についての研究も行う必要がある。また電子辞書の作成支援ソフトというものも現状では個々の辞書に依存して個別に作られており、辞書の共用ということを考えた場合にはソフトウェアの共用ということも考えなければならないであろう。

最近では共用を目指した辞書作成の動きも見られる。たとえばIPA（情報処理振興事業協会）の技術センタで開発された計算機用日本語基本動詞辞書は、汎用の辞書として共用を目指し、作成に当たっては言語学者や計算機科学者が協力して作業したものである。また通産省と郵政省の共同で設立された財基盤技術研究促進センタの下で計画されている電子化辞書のプロジェクトは、ICOT（財新世代コンピュータ技術開発機構）での電子化辞書の研究成果をもとに、やはり共同で辞書を開発しようというものである。このようにして辞書を開発する場合にはIPAのように計算機科学者と言語学者が協力し合って作業を進めるということが不可欠である。また、英日や日英といった二国間辞書を開発する場合にはかならずどちらの側にも複数のネイティブスピーカーを参加させることが必要である（バイリンガルの人がいればなおよいが、人材の確保が難しいであろう）。しかもそのネイティブスピーカーにしてもある程度言語学や計算機科学に素養のある人が望ましい（もちろん専門家であればなおよい）。しかも複数のネイティブスピーカーということを強調したのは、言語に対する感覚は個人個人で微妙に違っており、同じ文の解釈をめぐるもしばしば意見が対立することがあるからである。日本語についてもその人の生まれ育ちによって言語感覚には微妙な差があることが指摘されている。方言がからんでくるとこの差はもっと大きくなる。最も理想的な状態は両言語の複数のネイティブス

ピーカが常に同じ所で作業してお互いに生じた疑問についてディスカッションできるというものである。このようにすれば両言語間に生じた微妙なニュアンスの差、入出力の迅速な評価などが常にコミュニケーションを保ちながら行える。このような体制、環境作りも辞書そのものと直接のかかわりはないものの今後必ず問題になってくるであろう。同じ場所での作業が不可能な場合には何らかのコミュニケーション手段を確保し、疑問が生じた場合にはすぐに問い合わせが行えることが望ましい。このようなことを考えると問題は国際的なネットワークの構築といったところまで広がるが、ここではこれ以上の言及は避ける。

疑問が生じた時にすぐに問い合わせができるということは、上に述べた辞書作りの場面でのみならず、辞書を使用する場合にも必要なことである。これは辞書を作った後のサービス体制をどのようにするかという問題ともつながる。冊子体辞書の場合には改訂版が出されると読者は新しい版を買わなければならない。しかしながら本の場合には古い版でも使用する時にそれほど支障はないし、新しい版を買う時の負担もそれほど大きくはない。電子化辞書の場合には辞書そのものを既に改変していたり、辞書に付随するソフトウェアの蓄積があったりして改訂の程度によってはユーザに与える負担や損失が大きい。始めから共用を目指して十分に検討を加えればそのような問題は起きないという考え方もあるが、対象が自然言語である以上、時代とともに使用法、語彙などが変化すると考えた方が自然である。従ってたとえば改訂の際にはコンピュータの入れ替えの際によく行われるように、アップコンパティブルな辞書にする、すなわち従来の辞書を使用していた者に対しての影響を最小限に食い止めるようにする配慮が必要である。辞書そのものの改変がユーザに許されている場合には、ユーザの改変した部分と新たに改訂された部分とのマージングをとってユーザの負担をなるべく軽くすることが必要であろう。辞書作成のところで述べたのと同じように、辞書改訂の際にも改訂に伴う種々の問題を解決する支援ソフトウェアを作ることを考える必要がある。

以上のように自然言語コンピュータ解析用辞書の作成、改訂、運用に伴う種々の問題点について述べてきた。それぞれの項目についてざっと概観しただけであるが、どれ一つをとっても問題は大きい。またそれらの問題を解決するために払わねばならない努力や負担も大きいので、辞書を共用化することには大きな意義がある。共用化に伴う種々の問題点についても指摘した。これらの問題点を一つ一つどのように解決していったらよいかを考えていくことが今後の課題であろう。また実際に小規模な実験を行ってここに書き漏らした問題点を洗い出すことも必要になると思われる。それでもなお大規模な辞書を作成した時には大規模システム固有の問題点が出てくることがあるので、その点についても十分な検討が必要である。

3 自然言語解析における辞書の役割

3.1 自然言語処理系の構成要素

およそ自然言語を機械的に処理しようとするシステムにおいては、その言語に関する膨大な情報を何らかの形であらかじめ保持していることが必須である。これは、言語現象とそのアプリケーションに関わる骨組みを記述するアルゴリズム、比較的一般化しやすい現象の内容を記述する規則、および個別の事象に対するデータを記述する辞書という3種類の構成要素で整理することができる。これらの中でも、辞書、とくに「語」に関する辞書は冊子体形式での歴史をもち、我々に身近なものであるのみならず、自然言語処理のアルゴリズムのほとんどが辞書を通すことで始めて起動できるといった必須性から、とくに再検討を要しないほどの自明な要素とさえ考えられている。しかし、実際の自然言語処理システムにおいて、そこで辞書と呼ばれるものは、アルゴリズムや規則、さらには処理の目的や運用形態によって様々に規定をされており、その規定のされ方を明らかにしない限り、自然言語解析用辞書一般を見ることが、それらの間の共用可能性などについて議論することもできないであろう。

自然言語処理系の基盤としてのアルゴリズムはシステムに処理の理論的基盤を提供するものである。これは、規則や辞書にその基本モデルあるいは枠組みを与える。一つの理想としては、アルゴリズムが規則や辞書の作成基準を完結的に規定することが望まれる。しかし、これを完全に行うことが現実的に困難であるということもあって辞書や規則の自律性、独立性をいかにして保証するかということが、一方では重要である。アルゴリズムは規則や辞書の基礎をなすものとして普遍性をもつべきであるが、実際には異なる辞書の間の透視性としては逆転している場合が多い。このことは、異なる辞書の間で様々なレベルでの相互利用を図る上で注意すべき問題点である。

アルゴリズムは、また、即自的なシステム要求に対し、プログラムないしシステムとしての実現を保証するものである。自然言語処理系に対するシステム要求は、自然言語とその背後の世界の多様性を反映して、たびたび極めて多様であらゆるレベルにわたる。たとえば、自然言語を扱うための言語モデルの設定は、その言語モデル自体で自律的な諸々の要件を引き起こす。他方、自然言語処理系に対する各種の外部要件には、こうした言語モデルの中に取り込みにくい類のものも多い。そこで、現実のそうしたシステムを統一的なアルゴリズムの下に構築することには、しばしば、かなりの無理があり、そうした多様なレベルの多くの部分を辞書が吸収せざるを得ない場合も少なくない。

自然言語処理系における規則は、具体的な言語現象等に対する記述の集まりであるが、辞書がとくに個別を対象とするのに対し、より一般的現象を主に扱うものである。規則の扱う範囲はシステム

によって千差万別であるが、いずれにしろ、ある固定した言語処理のモデルの中における可変部分を受けもつと考えるとよいであろう。これは、プログラム・ロジックとして実現してしまう場合も多かったが、逆に、プログラムを抽象化して、システム固有のロジックは規則として記述できるような、いわば基本プログラム化をはかることも少なくない。

規則は辞書との境界において、より多くの問題をはらんでいる。辞書の各語彙項目の中には、色々の規則素性ないし規則そのものが記述される。また、その他の辞書の内容も、各種のカテゴリー分類や処理系のモデルあるいは言語モデルの上の表現に対する記述がほとんどである。これらは規則と密接に結びついているのみならず、しばしばそれらの境界は不明確である。

規則は、個々の語から離れた一般的な記述をするものであるけれども、現実の規則記述に際しては、辞書における特定の記述を期待したもの（あるいはその逆といってもよいもの）が少なくない。このことは、処理の内容が高度になるにつれて、ますます増大する傾向が見える。これは、辞書作成および保守の障害要因となるが、一方では辞書の共用化に弊害をもたらす。

辞書は自然言語処理系に対して、その処理のもととなるデータを供給する。自然文を入力として処理を行うシステムであれば、その各々の語を辞書の語彙項目に書き換えることにより、その後の処理が可能になる。辞書は概念としては、「語」だけでなく、様々の先験的なデータを収納しておく機能としても考えられるが、ここでは「語」を単位とするものとしている。自然言語処理系における辞書は、「語」をキーとして、各種規則を起動するためのパラメータとデータを処理系に供給する機能と定義することができる。

自然言語処理用の辞書は、各個別システムによって、様々な形式と内容で作成されている。そのもとづくアルゴリズムや規則（たとえば、文法理論と文法）によって異なるだけでなく、その表現能力を向上するために規則や手続きを導入するなどそれぞれの独自性を有している。このような中で、辞書の共用化を図るためには、広くアルゴリズムの共通化、普遍化の分析、さらにはその前提の下での規則の共用化、普遍化の分析を経ることが必須であろう。

3.2 自然言語解析の各フェーズにおける辞書の利用

(1) 形態素解析

形態素解析のフェーズは、少なくとも、入力文字列としての自然文の中で「語」の認定を行う。これは、自然文文字列を「語」の単位をもととした構造に置き換えることにより以後の解析を可能とする。ここでは、入力列での位置情報などの出現形態の情報とともに辞書の語彙項目相当の内容を与えることが主要な機能である。

語の同定のためには文字列レベルの照合を可能にする「語形」にあたるものを用いる。これは通常の辞書の見出しに対応するが、その他、語形の屈折や派生、接辞の扱いなどが必要であり、

これらはシステムにより扱いが異なる。また、とくに日本語の場合、語分割の処理が必要であるが、辞書の側でも、語単位のとり方の基準が問題であり、必ずしも一般的に定まっているとはいえない。このことは複合語など、欧米語でも発生する。

形態素解析においては、さらに文字列処理を含む比較的局所的なレベルでの様々の解析処理が加わることが多い。辞書未登録語に対しては辞書登録語相当の情報を推定処理により与えるための機能がある。また、辞書との照合に成功したものでも、その出現環境によって、辞書記述の内容への制限、変換あるいは合成を施す必要がある場合もある。たとえば、英語における「's」、ハイフン結合、派生語などを考えればよい。また、日本語における語分割の処理に際しても、単なる照合処理の範囲を越えたロジックが起動されるのが普通である。こうした処理のために、固有名詞、術語の構成語といった特殊な情報から、品詞レベルや一定の意味素性のような情報までが辞書情報として用いられる。ただし、ここで用いる情報やその方法論は、システムにより様々であり、これらを統一して扱うことは現在のところ、かなり困難であるといえる。

(2) 構文解析

構文解析というときに、そこで扱う範囲は、システムによって千差万別である。ここでは意味的な情報も含め、統語論的な操作を扱うフェーズを扱うことにする。この文法論的な処理を行うフェーズは、さらに、表層の構造の解析を行うものより深層に関わる解析の2レベルに分けることができる。

はじめのレベルは共通化がかなり可能に思われるもので、通常、文脈自由型文法相当の規則で表層構造を解釈する。特に、英語などのように統語論の拘束力が強い言語においてはこのレベルでの規則の表現能力は大きく、単なる品詞カテゴリーや性数格の呼応など、多くの規則をこのレベルに適用することができる。そのため、このような表層構造レベルでの文脈制限などの各種情報が辞書の語彙項目から採られ、用いられる。これは逆に、異なるシステムで辞書の内容が異なる要因ともなる。たとえば動詞文型の分類にしても色々なものがある。さらに、“tough”、“eager”といった例にもあるように、他の品詞についても同様の扱いをする場合もある。こうして文法の立場や処理の都合などから、辞書に要求する統語論の情報もまったく異なってくるというのが実態である。

これらの情報は、自然言語処理の上からは、相当にプリミティブなものであるが、冊子体形式のデータでも、現在入手できるものは、はなはだ不十分である。文法的背景を明確にした上で、構文解析処理に耐えうるだけの共通して使いうるデータが複数構築されるような状況が待たれる。

解析の次のレベルとしては、表層レベルの構造あるいは解析過程に対し、より深いレベルへの変形や構造作成操作、あるいはそのレベルにおける共起関係や構造の評価といった処理がある。その最も代表的な例は、格関係などにそった選択制限の操作である。選択制限などの情報は、表層レベルでの解析規則と異なり、通常は個々の語に対し与えられる。こうした記述は一般の冊子

体形式の辞書においても類似のものが見られるが、計算機で扱うためのものと人間のためのものとでは、やはり本質的ともいえる相違がある。これは、人の知識、経験、感性に訴えて理解を深めることを目指したものと、計算機上でのコンパクトなメカニズムのデータとして与えるものととの差異であり、自然言語処理用の辞書を考えるときに忘れるわけにはいかない点である。

選択制限は、ある語を軸として、把握された構造の上で何らかの関係をもった語としては許されないもの（あるいは必須なもの）を指定する記述である。制限先の個々の語をすべて記述することは不可能な場合が多い。そこで何らかのカテゴリーを設定することになる。実際、選択制限の定式化としては内在素性としての意味標識が主に用いられる。さらに、こうした辞書記述をサポートするために、選択制限処理との間に、カテゴリーないし語の間の関係を記述したという意味での余剰規則ないし、シンソーラスが用いられたりもする。

しかし、個々で様々な問題が生じる。たとえば、“kill”が“oneself”をとって「自殺する」の意味をもつか、“catch”が“cold”をとって「風邪をひく」という意味であるといった硬い規定ならば誰が記述してもさほどの差異ではないが、より抽象的な（それゆえ強力なものではある）カテゴリーに関しては、その記述の基準があいまいとなり、また、メカニズムとしても妥当性の評価がしにくいといった問題が発生する。また、“take”を「乗る」の意味でとるための「乗り物」といったグルーピングは、それ自体としては有効であるけれども、大量の語にたいしこれを行ってゆくと、莫大な数の、あるいは、互いに重なり合い、諸々の関係を互いにもつカテゴリーが生まれてしまう危険性がある。また、こうして与える内在素性としての意味標識が、語の意味記述に照応するのか否かという点も自明ではない。これが他システム間の関係となると、横並びも困難なほどに、内容のみならず、基準も異なっている。しかし、こうした選択制限の処理などは、現在ではある程度の普遍性を持つにいたっていると考えられるので、この種の言語現象を記述したマスター・データを実現し利用するための可能性と方法論についての模索が必要であろう。

その他、このフェーズでは、語に関する個別の各種情報が用いられる。たとえば、種々の様態の副詞とか、状態動詞であるといった情報など、比較的普遍性をもちうるものも多い。また、語の異なる品詞や語義の間での、その使い易さの基準や、用いられる分野の情報なども有効である。

構文解析処理のターゲットとしての構造表現はシステムによってまちまちである。これも個々のシステムにおける辞書の特殊性を強く規定している。実際、各システムの辞書の中核部分は、ほとんどこうした個別の構造に対応した記述に集中している。それが各辞書の最も価値があり、最もコストを要求する部分であるが、反面、他者の利用を困難にするものでもある。しかし、このレベルでの共有性による負担の軽減が最も要求されるべきものであることも事実であろう。

(3) 意味解析

ここでは意味表現の構造と、その上での種々の操作に関わる処理を意味解析と呼んでいる。こ

のレベルで既に実用化されているのは、具体的なアプリケーションに対応したセマンティックスとそれにもとづく具体的な操作に限られると言ってもよいであろう。普遍性をもった意味解析の枠組みは主に研究室段階にある。しかし、共有化しうる辞書を考えるとき、意味解析に利用しうる能力を残しておくことは重要である。実際、既存の冊子体の辞書などの情報の多くは、このレベルに結びつくべきものである。たとえば、辞書の意味記述の形式化を図ることは、その背景としての「世界」の先験的記述や推論規則の試行とともに、辞書作成者、自然言語処理系の開発者の双方にとって、実り多いものであろう。

(4) その他の利用

機械可読化されているものを含めて、冊子体形式の辞書の類は、自然言語処理系の開発において重要な位置をしめている。実際、そのようなデータは、システムの開発のための典拠データ、作業用資料として、あるいは事前・事後の分析用データとして、アルゴリズム開発、規則記述、辞書作成およびそれらの評価や保守の各フェーズにおいて用いられている。もちろん、そういったデータのそのままでの利用が可能な場合は、たとえば翻訳システムにおける専門語辞書の対訳のみの利用といったものに限られ、その場合でもかなりの手直しが必要となったりする。しかし、少なくとも、辞書エントリーの収集、頻度や重要度の分析、例文などのコーパス的な利用などが可能であり、さらに、辞書にある品詞・分野などのコードとしての転用や意味記述からの諸情報の抽出の実験なども試みられている。

また、自然言語処理への補助的な利用法としては、冊子体辞書型のデータの有機的な利用は必要である。たとえば、機械翻訳を例にとると、システムが直接用いる辞書とは別に利用者が翻訳前編集・後編集などのフェーズで、各種辞書、シソーラス、用語集、用例集さらには百科事典や専門知識を参照するニーズは高い。それらを有機的に利用できるトータル・システムの構築は、自然言語処理系の内部辞書とは別に、積極的に押し進めてゆくべきテーマであるだろう。

4 機械可読辞書の調査

4.1 調査項目

機械可読辞書について以下に示す調査項目を設定して、海外、国内における辞書事例の調査を実施した。

表 4-1 調査対象辞書に関する調査項目

- | |
|--|
| <ul style="list-style-type: none">① 各プロジェクトにおける機械可読辞書の種類、及びシステムの種類、及び各システムの機能概要（機能図）② 機械可読辞書のデータ項目、及び各データ項目の機能・目的（システムとの関連）③ 各データ項目のデータフォーマット（各データ項目の記述形式）④ 機械可読辞書のファイルフォーマット、及び各データ項目間の論理構造⑤ 現時点での単語収録件数、及び収録語の種類、及びそれぞれの目的⑥ 各データ項目の内容構成<ul style="list-style-type: none">【例】 ・見出し語に対する収録語の種類・品詞分類・意味マーカ、及びその体系・格フレーム、及びその体系⑦ 機械可読辞書作成上のデータエン트리・ルール（辞書作成マニュアル、及びサンプル）⑧ 機械可読辞書管理システムの機能概要（機能図）⑨ 機械可読辞書の作成、及びメンテナンス体制<ul style="list-style-type: none">・体制・方法・人数・資質・経費⑩ 他プロジェクト、及びシステムとの共用使用にあたり、辞書作成、及びメンテナンスを行っていく上での考慮すべき点⑪ システムの利用における利用者側の機械可読辞書の操作可能部分と操作方法<ul style="list-style-type: none">・利用者メンテナンス可能部分、システム側メンテナンス部分⑫ 提供可能性、及び提供方法⑬ 関連資料、参考文献 |
|--|

海外事例については、石川徹也専門委員会主査を調査員として派遣し、西独、ルクセンブルグ、オーストリア、イタリア、カナダの大学、研究機関等 9 機関を調査した。詳細は、4.2 節、4.3 節に述べられている。

国内事例については、アンケート（26 事例を対象とした）およびヒアリング（5 機関を対象とした）により調査を実施した。詳細は、4.4 節、4.5 節に述べられている。

ただし、海外調査については時間的制約、国内調査については、対象機関の機密事項に触れる項目があるために表 4-1 の調査項目のすべてについて明らかにできたとは言えないが、国内外の事例の現状のおよその姿は把握しえたと考えられる。

4.2 海外調査概要

機械可読辞書（MRD）の海外調査は、昭和 60 年 10 月 13 日（土）から 10 月 27 日（日）に、表 4-2 に示す機関を訪問、調査を行った。以下、順にその概要を示す。

西ドイツの訪問機関は、GID（西ドイツ情報ドキュメンテーション研究所）の国際部長 Mönch 氏のアレンジにより、当地にて予定を設定した。そのためもあって、日程上、各機関とも表敬訪問の形になってしまった。

ザールランド大学、シュトゥットガルト大学、EC 翻訳センター（ECAT）は、自動翻訳システムについての話が中心で、ボン大学、西ドイツ言語研究所は、開発中の MRD、およびテキスト・データベースの概観について、ピサ大学は MRD のイタリア、および EC のプロジェクトをいかに考えていくかということ、国際用語学情報センター（INFOTERM）は、TERMNET（International Network for Terminology）の必要性・意義について、ウォータルー大学は the new OED の機械可読化の開発について伺うことができた。しかし、日程の関係で、どれも概要にとどまり、当初の詳細なる調査項目を、完全に把握できたとは言えない。

表 4 - 2 自然言語解析用辞書 海外調査 訪問機関一覧

№	機 関 名	訪 問 相 手	国
1.	西ドイツ情報ドキュメンテーション 研究所 G I D (Gesellschaft für Information und Doku- mentation MBH)	Dieter Monch (国際部 部長)	西ドイツ
2.	ボン大学 コミュニケーション・音 声研究所 I K P (Institute für Kommunikation sfors- chung und Phonetik der Universität Bonn)	Winfried Lenders (I K P 所長, 言語学部教授) Jan Brustkern (I K P 研究員, Lenders 教授の助手)	西ドイツ
3.	ザールランド大学 Universität des Saarlandes	Harald H. Zimmermann (情報科学科 教授) Krishna Sharma (情報科学科 教授) Edith Kroupa (情報科学科 博士課程 女子学生)	西ドイツ
4.	シュトゥットガルト大学 Univesität Stuttgart, Institute für Informatik	羽中田 賢治 教授 (情報研究所)	西ドイツ
5.	西ドイツ言語研究所 IDS (The Institute für Deutsche Sprache)	Wolfgang Teubert (G I D 副所長) Klaus Wothke (研究員) 在 間 進 (東京外国語大学 ドイツ語学科 助教授)	西ドイツ
6.	EC 翻訳センター E C A T (European Center for Automatic Translation)	Michel Lebrun (ECAT マネジャー) Cay - Holger Stoll (マネージング ディレクター)	ルクセン ブルグ
7.	国際用語学情報センター INFOTERM (Internatio- nal Information Center for Terminology)	Christian Galinski (副所長)	オースト リア
8.	ピサ大学 University of Pisa, Department of Linguistic	Nicoletta Calzolari (言語学科 助教授) Antonio Zampolli (言語学科 教授)	イタリア
9.	ウォータルー大学 University of Waterloo, Center for the New Oxford English Dictionary	Gayle Johannesen (事務長 女性) J. Stubbs (Univ. Oxford Press の OED 編集部からの出向) F. W. Tompa (計算機学科 助教授)	カナダ

4.3 海外調査結果

4.3.1 西ドイツ情報ドキュメンテーション研究所 (G I D)

G I D (Gesellschaft für Information und Dokumentation MBH)

住 所 : Ahrstrabe 45, D-5300 Bonn 2, (Bad-Godesberg),
Federal Republic of Germany

電 話 : 西ドイツ -0228-373326

日 時 : 10/14 (月) 10:30 ~ 12:30

訪問相手 : Dieter Mönch (国際部 部長)

内容:

G I Dの国際部長D. Mönch 氏によると、西ドイツにおいては、国レベルの、また自然言語処理用の統一的な機械可読辞書(以下、MRD: Machine Readable Dictionary とする)の開発プロジェクトは現在のところないとのこと。西ドイツにおいては、主に、自動翻訳システムの研究・開発プロジェクトに必要とするMRDがプロジェクト単位で作成されているとのこと。そこで、国内の主に自動翻訳プロジェクトを推進している大学を訪問してほしいということで、訪問スケジュールの調整を受けた。以下、簡単にG I Dの機関の機能について紹介する。

G I D (英語名: Society for Information and Documentation) 自身は西ドイツ政府の情報政策を具体的にプランニングする機関である。G I Dは、1971年5月に、西ドイツ内務省が、西ドイツの情報システムのあるべき姿について展望を発表、すなわち、「分野別情報センターおよびその情報ネットワーク構想」に基づき、西ドイツ研究技術省(the Federal Ministry of Research and Technology)が1974年に策定した「情報・ドキュメンテーション推進計画」に基づき、1977年に設立、1978年1月1日から活動を開始した政府機関である。資金は、西ドイツ政府が65%, 州政府が35%出資している。

ボンに本部、フランクフルトに支部があり、また東京、ワシントン(USA)に支所があり、現在約280名の職員が働いている。主に、現在西ドイツに存在する16主題領域の情報センターと特許、公害等共通主題領域の4つの情報センターに対する年間20~30億円の情報政策費を管理している。主たる活動内容は、以下の通りである。

- 1) 情報システムに関する研究・開発
- 2) 技術サービス
- 3) 専門知識の提供および専門家の派遣
- 4) 情報科学および情報活動に関する情報サービス

5) 情報およびドキュメンテーション活動機関およびG I Dのメンバー機関へのコンサルティング・サービス

6) 州および連邦政府の研究プロジェクトの管理

7) 国内および国外機関との調整窓口

現在、日独科学技術政府間協定に基づく、情報科学分野のプロジェクトは、下記2点が推進されているとのこと。

1) 日本の科学技術情報の西ドイツでの活用を意図し、特殊法人日本科学技術情報センター

(J I C S T : Japan Information Center for Science and Technology) で作成されているデータベースサービス J O I S (J I C S T Online Information Service) のオンライン検索の普及。

2) 上記情報サービスの普及を意図し、日独自動翻訳システムの研究開発。

なお、Mönch 氏がアレンジして下さっていたが、日程の関係で訪問ができなかった下記2人の方はMönch 氏の所に事前に資料の提供を下された。

1) 江沢健之助(テュービンゲン大学独文科準教授)

先生の研究の著“Sprachsystem und Sprachnorm”(言語の体系と語の通常態)、Max Niemeyer 社から出版、を基に、「新しい形」の「和独辞典」の作成、そのデータベース化を推進されているとのこと。

(拝受資料)

・(海外塾員日より)テュービンゲン大学昨今、三田評論、1985年8月8日号、p.98~102

2) Burghard Rieger (the Technical University of Aachen, ドイツ語学部教授)

上記大学・学部では、西ドイツのthe NEW Ministry of Science and Researchの支援により a project in Computational Semantics of the MESY (Arbeitsgruppe f. mathem. - empirische Systemforschung)を推進。当研究は、談語入力データから意味レベルを自動的にFuzzy関数化、さらに知識レベルを生成することを目的としている。

(拝受資料)

・Riger, B.

Lexical relevance and semantic disposition—On stereotype word meaning representation in procedural semantics—, p.388~400, in Meaning and the Lexicon, Pieter A. M. Seuren (ed.), Foris Publications, 512p., 1985

GESELLSCHAFT FÜR INFORMATION UND DOKUMENTATION MBH (GID)

Organizational Structure
01.05.1984

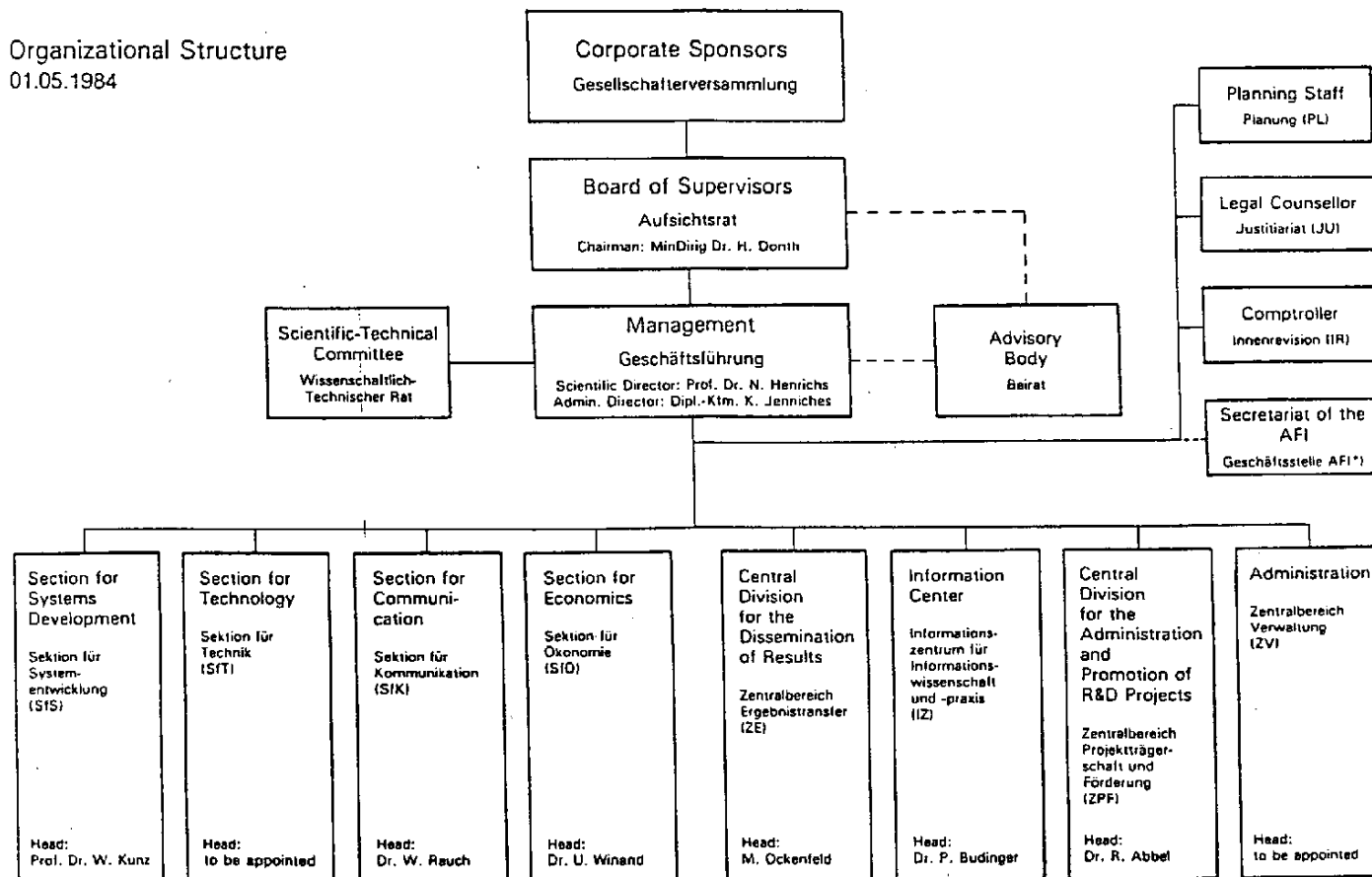


図 4-1 G I D の 組 織

*) AFI = Arbeitsgemeinschaft Fachinformation (Association of Specialized Information Providers)

・ Rieger, B. B.

Fuzzy Representation Systems in Linguistic Semantics : An
Empirical Approach to the Reconstruction of Lexical Meanings
from East - and West - German Newspapers

Progress in Cybernetics and System Research, Vol. XI, p. 249~
256, 1982

・ Rieger, B. B.

Semantic Relevance and Aspect Dependency in a Given Subject
Domain - Contents driven algorithmic processing of Fuzzy
word meanings to form dynamic stereotype representations —

Proceedings of COLING 84, p. 298~301, 1984

・ Rieger, B. B.

Generating Dependency Structures of Fuzzy Word Meanings in
Semantic Space

Proceedings of the XIIIth International Congress of
Linguists, p. 543~548, 1983

4.3.2 ボン大学 コミュニケーション・音声研究所 (IKP)

IKP (Institut für Kommunikationsforschung und Phonetik
der Universität Bonn)

住 所 : Poppelsdorfer Allee 47, 5300 Bonn 1,
Federal Republic of Germany

電 話 : 西ドイツ -0228-735638

日 時 : 10/14 (月) 13:30 ~ 14:30

訪問相手 : Winfried Lenders (IKP 所長, 言語学部教授)

Jan Brustkern (IKP 研究員, Lenders 教授の助手)

内容:

Lenders 教授から上記 IKP (英語名: The Institute for Communication
Research and Phonetics of University of Bonn) の研究目的は, 次の 3 点にあるこ
とを伺った。

- ① ドイツ語の言語学研究 (主に音声処理)
- ② 機械翻訳の研究・開発

③ コミュニケーションに関する音声・音韻学の研究（主に辞書の構築）

次に、Brustkern研究員に、“the German Word - Data - Base”の検索を見せてもらった。当MRDは、西ドイツにおいて、自動翻訳、自然言語処理、情報検索、テキスト処理等の研究・開発プロジェクトにおいて作成されている各プロジェクト用MRDを一本化し、それを利用できるように整備している上記IKPのプロジェクト（プロジェクト名：AKU）のものである。

当MRDは、ドイツ語の基本見出し語（語幹：stem）データベースを目的とし、とくにドイツ語特有の複合化、および屈折語（活用語尾）の研究に使用することを目的としている。

収録語数 300,000語	・ 語幹（stem）15万語、ただし基本語（base word） (EX) : Haus ・ 上記語幹に対し、複合化語尾、および活用語尾 (屈折語：inflected wordを15万語) (EX) : ~alter 家庭用祭壇 ~andacht 家庭での礼拝 ~anzug 部屋着
------------------	---

図4-2は、1983年7月現在において収録されているプロジェクト別の辞書リスト、図4-3は、一本化し利用するシステム（システム名：ALEXSYS）の機能図、図4-4は、辞書データ項目である。

このプロジェクトはECの翻訳支援用語辞書（EURODICAUTOM：ターミノロジーデータベース）との連携を図っているとのこと。

（資料1）英文

Brustkern, J., Schlimgen, E., & Schulze, W. "Two Papers on the creation of the German word-data-base/Bibliography on MRD."

Name	Size	Origin	Name	Size	Origin
CONDOR	843	Dictionary of function words used in a system for automatic analysis and indexing of natural language documents; SIEMENS AG München	MOLEX	24000	Dictionary of the question-answering system FLIDIS; 1972-79 IDS Mannheim
LIMAS		Research project "Linguistische und maschinelle Sprachverarbeitung" 1965-1976, IKP Bonn	SADAW	110000	Dictionary for automatic text analysis and machine translation developed at the SFB 100 Saarbrücken since 1976
	130000	list of all word forms in the LIMAS-Corpus	SDW	100000	former version of the SADAW dictionary; 1970-75
	1160	list of German prepositions	VALEX	1600	Dictionary of verbal valences compiled by U. Engel and H. Schumacher; 1976 IDS Mannheim
	1427	nouns with prepositions	WAHRIG	4000 (only verbs)	Investigation on the structure of natural languages. Research project of G. Wahrig at the University of Mainz; 1974-78
	5145	verbs with prepositions			
MACKENSEN	117000	machine readable version of the dictionary of Standard Modern German by Lutz Mackensen compiled by G. Kandler and K.D. Bunting; IKP Bonn			

Table 1

図 4-2 IKPにおける機械可読辞書のリスト

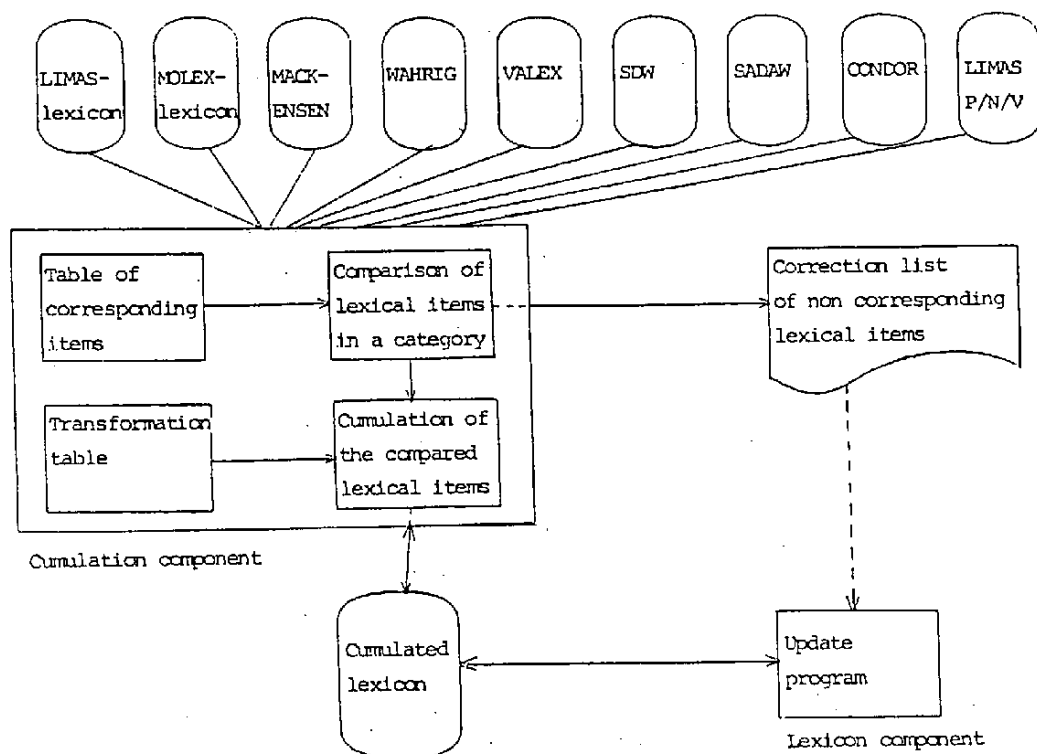


図 4-3 ALEXSYSシステムの機能

THE GENERAL STRUCTURE OF A MACHINE- READABLE DICTIONARY

1. WORD-ENTRY

The information category 'word-entry' is divided into two subcategories: 'word-form' (graphical representation) and 'versions' (including informations about orthographical syllablesplit and other variants).

2. PHONETICS/PHONOLOGY

Including different types of transcription and hints on prosodic features.

3. WORD-CLASS

The informations are based on a special classificational system.

4. MORPHOLOGY

Specification of basic forms and full forms, lemma-name of full forms, notes on both inflectional and word-building morphology.

5. SYNTACTICAL FEATURES

Informations on surface and deep structures according to different models of grammar.

6. SEMANTIC FEATURES

Informations on the description and classification of the meaning of the word-form, for example 'semantic markers', semantic relations, surroundings, definitions.

7. PRAGMATIC FEATURES

Notes on the use of the word-form: quotations, frequency, special field usage.

図 4-4 "the German Word-Data-Base" の辞書データ項目

4.3.3 ザールランド大学

Universität des Saarlandes

住 所 : Im Stadtwald, 6600 Saarbrücken,
Federal Republic of Germany

電 話 : 西ドイツ -0681-3023542

日 時 : 10/14 (月) 18:30 ~ 21:00

訪問相手 : Harald H. Zimmermann (情報科学科 教授)

Krishna Sharma (情報科学科 教授)

Edith Kroupa (情報科学科 博士課程 女子学生)

内容 :

○ Zimmermann 教授には、挨拶のみ。

Sharma 教授から、当学科のプロジェクトである自動翻訳システムの位置づけを伺い、デモを見る。

1) 日独政府間科学技術協定に基づき、情報科学分野 (日本側代表 京大 長尾教授, 西ドイツ側事務局 G I D) において、現在、日独自動翻訳プロジェクトを推進している。

当プロジェクトには、2つのプロジェクトがある。

ザールランド大学は、その1つで (プロジェクト名 SUSY-DJT) 京大 長尾研究室の自動翻訳システムによる、論文タイトルの日英翻訳結果を受け、ザールランド大学情報科学科の自動翻訳システム (SUSY) において英独翻訳を行う。この結果として、日独の翻訳を可能とする。

当プロジェクトは、1983年5月から始まり、今年で終わる。

今後についてどうするかは、未定とのこと。

図4-5は、“Ishikawa visited the University at the Monday.”という独文を、その場で入力、独→英翻訳デモンストレーションしてもらった画面のサンプル・コピーである。途中の詳細な操作は省略するが、すべてオンライン操作により逐文訳を可能としている。

(資料2) 独文

Projekt SUSY-DJT: Deutsch - Japanische Titelübersetzung mit
English als Switching-language

なお、当プロジェクトの開始にあたり、テキスト (文献) 検索システムを開発したとのこと。

独文入力

SATZ NR. 1 ENTHAELT FOLGENDE SEGMENTE :
1. SEGMENT (NR. 1) ***** OBERSTER KNOTEN *****
WORTLAUT : Ishikawa besuchte die Universitaet am Montag
IST HAUPTSATZ
DIE DRUCKPARAMETER LAUTEN :
DATEI 17-DAE SEEICHER 14 MODUS 1 ISVSCH 1



未知語及び品詞情報付与

SNR	WNR	TEXTWORTFORK	WKL	LEMMANAME	STW
1	1	Ishikawa	EIG	ISHIKAWA	SUB
1	2	besuchte	FIV	BESUCHER	VRB
1	3	die	ARTR	D-(ARTB)	FWK
1	4	Universitaet	SUB	UNIVERSITAET	SUB
1	5	am	PRP	AN(BAT)	FWK
1	6	Montag	SUB	MONTAG	SUB
1	7	*		*	



構文解析結果

Ishikawa <ISHIKAWA * EIG > besuchte <BESUCHER * ABJ > die<
D-(ARTB)* ARTB > Universitaet<UNIVERSITAET * SUB>am <AN(DAT)*
PRP > Montag <MONTAS * SUB > *



翻訳結果出力

Ishikawa besuchte die Universitaet am ISHIKAWA VISITTED THE
UNIVERSITY ON
Montas * THE MONDAY *

ISHIKAWA VISITTED THE
UNIVERSITY AT
THE MONDAY *

図4-5 SUSYにおける独和翻訳例

(資料3) 英文

The Text Database SUSYDJT-Short description and directions for use - (Version 1.0)

- 2) 上記1) のプロジェクトとは別に、ザールランド大学では1985年より西ドイツ調査・研究技術省 (the Federal Ministry of Research and Technology) の支援により、独英自動翻訳システムの開発を開始した。

当システム名は、ITS (Informative Translation System) といい、知識情報を活用するシステムを目指す。

当面、西ドイツ特許局 (the German Patent Office)、あるいは、技術情報センター (the Specialized Information Center Technik - FIZ 16-) の情報の翻訳を目的としていく。

なお、Zimmermann教授は、ECの翻訳プロジェクトであるEUROTRAの西ドイツ側代表である。

○ Manfred Thiel (電気工学科 言語研究室)

10/15 (火) 10:30 ~ 12:00 (午後Luxemburgへ移動)

Zimmermann, Sharma教授は、大学院の試験のため、Thiel助手からSUSYについて聞く。

SUSYでは、現在 ロシア語→独、仏↔独、英↔独、エスペラント語↔独の開発を推進している。

辞書は、独 ↔ 英の場合

- ①語幹辞書 (stem dictionary) (約150,000語)
- ②意味解析用辞書 (semantic analysis dictionary)
- ③形態生成用辞書 (morphological production dictionary)

の3種に分け、開発を行っている。

なお、SUSYプロジェクトで現在維持しているMRDの内容 (データベース名: SADAN) を図4-6に示す。SADANを管理するシステム (システム名: Deutsches Analyze - Lexikon) により、図4-7のような辞書内容のメンテナンスができる。

(資料4) 独文

Krebs, P. "DAS SUSY-LEXIKONSYSTEM Technical Report - 1985年8月"

なお、テキスト検索システムをはじめ、SUSYシステムは、Siemensのコンピュータ (OSはBS2000) を使用。検索ソフトウェア (DBMS) はGOLEMを使用とのこと。

Wörterbücher und Anzahl der Einträge

Analysewörterbücher

syntaktische Analyse

Deutsch	143.000
Englisch	10.000
Französisch	50.200
Russisch	15.000

semantische Analyse

Deutsch	75.000
Englisch	5.400
Französisch	1.700
Russisch	2.700

Übersetzungswörterbücher

Russisch - Deutsch	9.000
Französisch - Deutsch	9.000
Deutsch - Französisch	5.500
Englisch - Deutsch	12.900
Deutsch - Englisch	7.300
Esperanto - Deutsch	4.000
Deutsch - Esperanto	4.000

Synthesewörterbücher

Deutsch	14.500
Englisch	2.600
Französisch	8.800

図 4-6 辞書データベース S A D A N

```

*****
*   Deutsches Analyse-Lexikon
*   SADAN
*****
* Lexikon erstellt am : 250681
* letzte Änderung am : 101085
* Lexikon enthaelt : 142947 Eintraege
* auf : 4763 Datenseiten
* mit : 62 Indexseiten
*****

```

```

WL1: HAUS SDW
WL2: HAUSEN 110681
IBED = 1
RECTYP = 1
Z 1= 0 7 0 0
Z 5= 0 29 0 0
Z 6= 0 1 1 0

```

```

Verbstamm HAUS
Infinitiv HAUSEN
Z6/3 enth. 1 Infinitivstamm (INEI)
Z1/2 = 7 Praesenskonjugation wie: nisse
Z2/2 = 0 Infinitiv endet auf -en
Z6/2 = 1 Imperativ aus diesem Stamm bildbar
*** SYNTAKTISCHE INFORMATIONEN ***
Z3/2 = 0 persönlich gebrauchtes Verb
Z3/3 = 0 Reflexivphänomen unmöglich
Z5/2 = 29 mögl. Praep. Objektiv. Opt
Z9/1 = 0 Korrelat. nicht möglich

```

```

WL1: HAUS LUCKHARDT
WL2: 120384
IBED = 1
RECTYP = 2
Z 1= 3 0 4 0
Z 2= 4 1 0 1
Z 3= 0 29 0 0
Z 4= 0 0 0 2

```

```

WL1: HAUS LUCKHARDT
WL2: 120384
IBED = 1, RECTYP = 2
Z1/1 = 3, Z2/4 = 1 Substantiv neutrum Singular
Z2/1 = 4 Genitivendung -st -ES
Z2/2 = 1 Dativendung -st (-E)
*** SYNTAKTISCHE INFORMATIONEN ***
Z3/2 = 29 Praep. Valenz -st -Ont
Z4/4 = 2 Subst. kann in FS-Typ 2 vorkommen

```

```

WL1: HAUS KREBS
WL2: 10880
IBED = 9
RECTYP = 32
O(1): ( 0 ) jwk = 40, jstw = 0, jbed = 9
O(15): ksubk = 1, kgrph = 1, kklal = 0, kklal2 = 0
O(16): kklb(1) = 19, kklb(2) = 0, kklb(3) = 0, kklb(4) = 0

```

```

WL1: HAUS KREBS
WL2: 10880
IBED = 9
RECTYP = 32
O(1): ( 0 ) jwk = 40, jstw = 0, jbed = 9
O(15): ksubk = 1, kgrph = 1, kklal = 0, kklal2 = 0
O(16): kklb(1) = 19, kklb(2) = 0, kklb(3) = 0, kklb(4) = 0

```

```

WL1: HAUSARBEIT SDW
WL2: 50381
IBED = 1
RECTYP = 2
Z 5= 0 0 8 0

```

```

WL1: HAUSARBEIT SDW
WL2: 50381
IBED = 1, RECTYP = 2
Z1/1 = 0 Z2/4 = 0 Substantiv feminin Sg. ZEL
Z2/1 = 0 Genitivendung -st
Z1/3 = 0 Nominativ Plural endet auf -EN

```

図 4 - 7 SADANの辞書内容メンテナンス

4.3.4 シュトゥットガルト大学

Universität Stuttgart, Institut für Informatik

住 所 : Azenbergstr. 12, Herdweg 51, 7000 Stuttgart 1,
Federal Republic of Germany

電 話 : 西ドイツ-0711-2078

日 時 : 10/18 (金) 10:50 ~ 12:10

訪問相手 : 羽中田 賢治 教授 (情報研究所)

内容:

ザールランド大学と同様、日独政府間科学技術協定による一貫として日独自動翻訳の実験・研究を推進 (システム名: SEMSYN) している。当大学のプロジェクトは、ザールランド大学とは異なり、富士通との契約により ATLAS II による日本語文の意味解析を含む最終の解析結果をデータとして受け取り、当大学において直接独語文に変換生成する実験・研究である。実験研究期間は1983年12月から3年間であり、1985年の12月をもって終了の予定である。以後どうするかについては、日独科学技術協定の推進の問題でもあり、また富士通側の考え方との調整もあり、現時点では何も決っていない。機械可読辞書は、現実験・研究の範囲で持っている。当大学の実験研究は、意味解析までも含むもので、あくまでもアルゴリズム的な研究を中心としているので、大きな辞書は現時点では維持していない。

(資料5) 英文

Laubsch, J., Rosner, D., Hanakata, K., & Lesniewski, A.

"Language Generation from Conceptual Structure: Synthesis of
German in a Japanese/German MT Project" COLING-84
Proceedings.

(資料6) 独文

Rosner, D. "Generierung von Deutsch aus einer Semantische

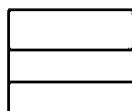
Representation - Zu Stand und Perspektiven des Projekts SEMSYN"

なお、プロジェクトの概要について、日本国内で紹介されたときのOHP資料 (図4-8~図4-12) をもって説明いただいた。

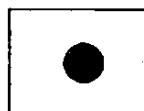
(図4-8) 日独科学技術協定に基づく日独自動翻訳実験プロジェクトの概要

(図4-9, 4-10, 4-11) SEMSYNプロジェクトの概要

(図4-12) 日独自動翻訳実験プロジェクト完了後に想像される運用形態



FRG



JPN

Agencies : BMFT/GID

STA/JICST

coordinator: Mönch/GID-IB

Shoji/STA

contact

person : Zimmermann/
(comp.ling) Univ. Saarbrücken

Nagao/Univ.Kyoto

Projects : SUSY-DJT (German↔English) (English↔Japanese) EJT

Univ.Saarbrücken,SFB100

Univ.Kyoto/JICST/ETI

/FUJITSU,HITACHI,

NEC

SEMSYN (German←Sem.rep) (Sem.rep←Japanese)

Univ.Stuttgart

ATLAS/U

FUJITSU.Lab.

TargetDB Flzen, TIB,DIMDI

JOIS,JAPATIC,RIPS

PATENT

NEEDS

Network

(LAN) — DATEX-P — VENUS-P — DDX — (LAN)

図4-8 Cooperation between FRG and JPN in the field of
information & Documentation (I & D)

Projekt

S E M S Y N

Ausführende Stelle : Universität Stuttgart Institut für Informatik

Projektträger/geber : GID/BMFT

Ziel : Generierung v.deutschen Titeln aus der
Semantischen Repräsentation Japanischer TitelZusammen arbeit : Fujitsu Laboratories/Kawasaki,JAPAN
(ATLAS/U)

Mitwirkende : Hanakata,Laubsch,Lesniewski,Roesner. + stdten

Rechnerumge : LISP-Machine SYMBOLICS 3600,
VAX-11/780-Anschluss (4 Terminals)

Projektdauer : 1.7.'83 - 30.6.'85.

Gesamtausg. : ~650KSM

図4-9 SEMSYNプロジェクトの概要

Weitere für SEMSYN-Fortsetzung relevante Aufgaben :

On-line Experiment

Pilotstudie zum Zugriff auf eine japanische Datenbank über internationale und nationale Datennetze

(Siehe Sitzungsprotokolle am 26.7.83

Protokoll vom 5. th. German - Japanese Meeting in Ia.D.
am 25.26.27.Okt.1983)

Simulation :

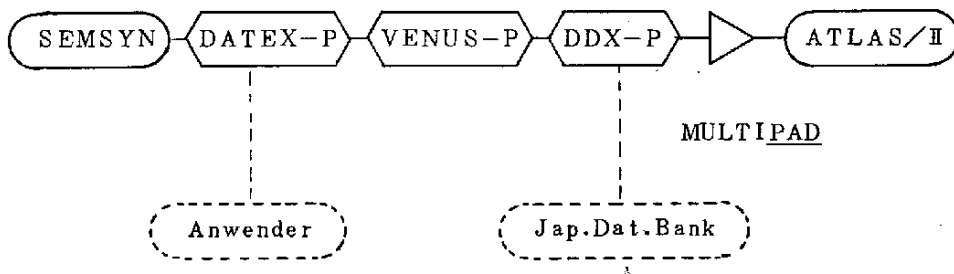


図 4-10 SEMSYNプロジェクトの概要

SEMSYN GOAL

Generation of 3000 German doc. titles in the field of "information processing" from the semantic representations of Jap. titles produced by ATLAS/U of FUJITSU Lab.

Also experiment with abstracts.

MEMBER : Hanakata, Lanbsch, Lesniewski, Roesne and Students

Hard.resource : Symbolics 3600, VAX-11/780 (4 Terminals)

Period : July 1.83 - June 30, 85.

図 4-11 SEMSYNプロジェクトの概要 (目標)

Long term goal G-J Cooperation in Information & Documentation

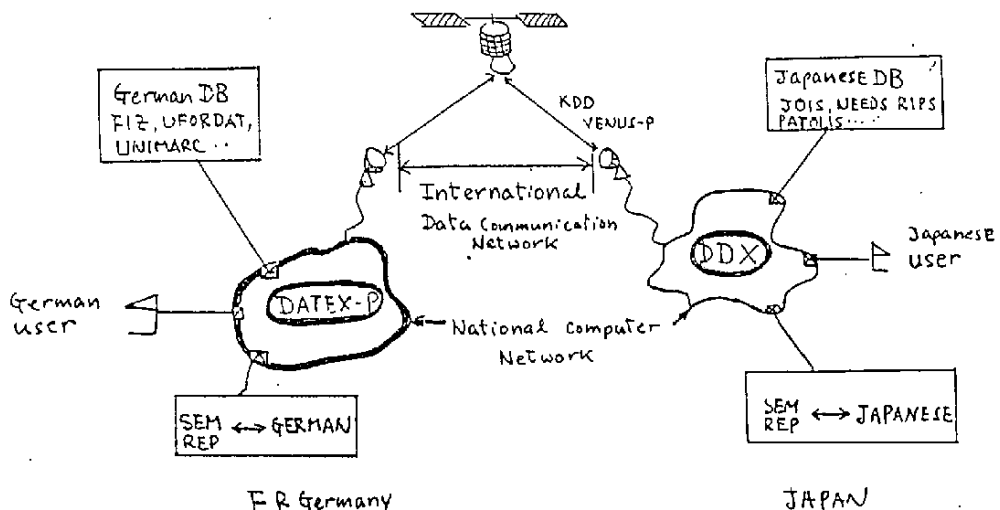


図 4-12 日独自動翻訳実験プロジェクト完了後に想像される運用形態

4.3.5 西ドイツ言語研究所

IDS (Institut für Deutsche Sprache)

住 所 : Friedrich - Karl - Str. 12, 6800 Mannheim 1,
Federal Republic of Germany

電 話 : 西ドイツ 0621 - 4401

日 時 : 10/18 (金) 15:00 ~ 17:00

訪問相手 : Wolfgang Teubert (GID 副所長)

Klaus Wothke (研究員)

在 間 進 (東京外国語大学 ドイツ語学科 助教授)

内容:

IDS は、日本の国立国語研究所に当る国立の機関で、ドイツ語の語彙調査、言語研究・言語傾向を研究・調査することを目的としており、とくに目的とする応用システムの開発は行っていない。そこで、言語研究のために、1983年から以下3つのデータベースプロダクション(プロジェクト名: LEDA: a Lexicographical Data Base for German)を行っている。使用計算機は Siemens. DBMS は IBM の STAIRS ライクの REFER である。

- ① フルテキストデータベース (Text Bank) … 例として、新聞・雑誌・学校の教科書・等
現在約 100 万件

② 語幹および活用語辞書 (Dictionary Bank)

③ 解析結果のデータバンク (Result Bank)

LEDAによるデータは、とくに、ドイツ語辞書の作成において辞書編成を支援するのに役立つものと考えている。

西ドイツにおける国レベルの言語解析用辞書の開発は行われていない。Brockhaus という出版社では、自社において冊子体辞書の機械可読化を図り、当出版社の辞書作成に使用している。研究用としても、著作権の問題、版権維持の問題があり、広く利用に供されていない (未公開)。

西ドイツ全体では、言語解析用辞書の作成としては、ボン大学の Lenders 教授の所の作成が唯一であり、ECの翻訳支援のためのオンラインサービスである辞書EURODICAUTOMとの関連を図っている。それについて、まとめた下記の本があり、参考になると思う。

JIPDECのプロジェクトとして参考になるのは、SIEMENSのプロジェクトであると考ええる。ただし、ECの次期自動翻訳システムの開発プロジェクトであるEUROTRAの西ドイツ側のスタッフは、ザールブルッケン大学のZimmerman教授グループの所である。

(参考図書) 独文

Heß, K., Brustkern, J., & Lenders, W. "Maschinenlesbare Deutsche Wörterbücher" Niemeyer Verlag, Tübingen, 1983.

4.3.6 EC 翻訳センター (ECAT)

ECAT (European Center for Automatic Translation)

住 所 : 9-11, rue Jean-Pierre Sauvage, L-2514,
Luxembourg

電 話 : ルクセンブルグ -00352 - 438388

日 時 : 10/16 (水) 13:30 ~ 15:00, 18:30 ~ 20:00

訪問相手 : Michel Lebrun (ECAT マネジャー)

Cay-Holger Stoll (マネージング ディレクター)

内容:

ECATは、ECの要請によりSYSTRAN(株)50%とルクセンブルグの大手製鉄会社ARBEDの情報処理会社グループ (Info ARBED) 50%の出資により1984年に設立された自動翻訳システムSYSTRANによるECの自動翻訳サービス・センターである (図4-13)。

現在、サービス・センターとしての整備、顧客営業、サービス体系の整備を行っている。勿論、SYSTRANの単独販売活動も行う。

ECは、現在も、"Multilingual & Multipurpose information processing" を

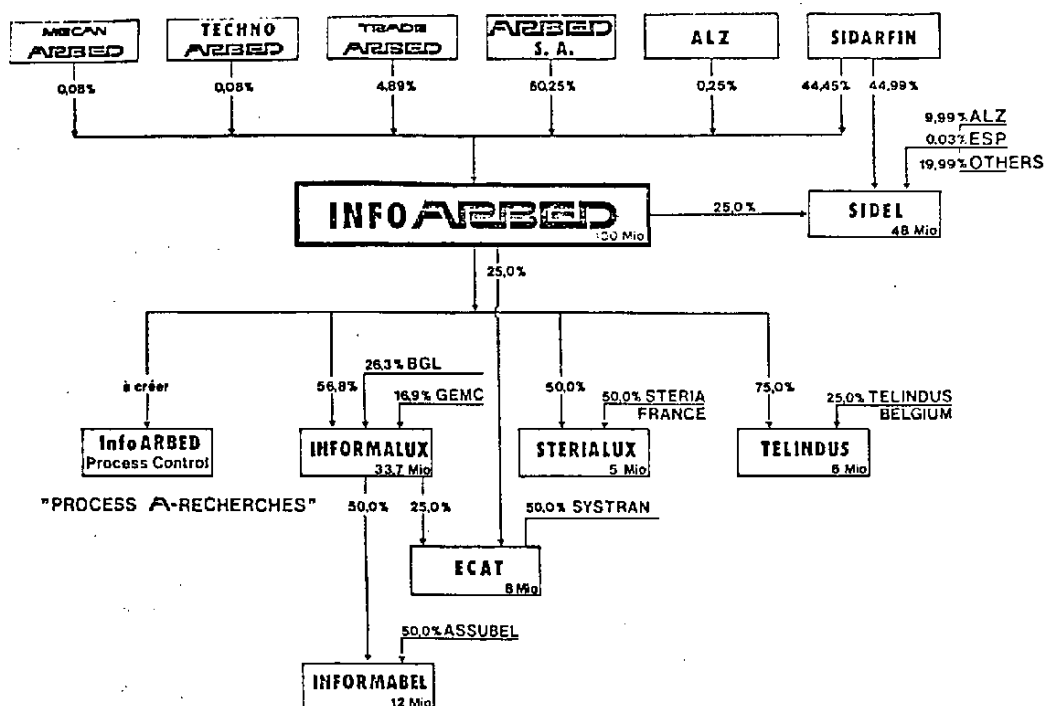


図 4-13 ECAT の組織化構成図

整備・運用している。その一つにEURODOCというオンライン情報検索システムにより情報提供サービスを行っている。自動翻訳プロジェクトはその一つのプロジェクトとして位置づけられる。その自動翻訳サービスを、この度のECATを通し行おうとするものである。すでに、パリにあるセンターを通し、英→仏、仏→英、英→独、英→伊、仏→独（ただし辞書は未完）のサービスを実施開始している。

MRDは、下記のように同じく整備中である。

- ① 独 → 英 約150,000語
- ② 英 → 仏 約120,000語
- ③ 仏 → 英 約120,000語
- ④ 英 → 独 約100,000語
- ⑤ 英 → 伊 約120,000語
- ⑥ 英 → アラビア語 約180,000語
- ⑦ 仏 → 独 開発中

データ項目は、大別すると語幹、活用情報、意味情報である。SYSTRANは、EUROTRAとは一線を画し、すでに実用レベルにある。EUROTRAは、現在、机上設計レベルとみている。

E C A Tの役割りは、いかに多くのユーザをふやし、利用していただくかということにある。また、現時点では、EURODICAUTOMを、たとえばデータ項目のちがいがらみても利用する計画はないとのこと。

(資料7) SYSTRANパンフレット

(資料8) Advantages of Machine Translation

(資料9) General System Specification for the Application of the SYSTRAN System.

ただし、SYSTRANシステムの辞書の基本的な整備が完了した時点で、より語数の必要性を感じた時は、EURODICAUTOMを語幹辞書(stem dictionary)に取込むことを検討するかもしれない。また、SYSTRANを利用したいユーザがEURODICAUTOMを翻訳作業の支援ツールとして利用することは大いに有効であると考えてるので、SYSTRANによるサービスが充分なされるようになってもEURODICAUTOMの存在を否定することにはならないと考えている。

E C A Tを組織化し、翻訳サービスを開始しているのに、EC内で現在多額の予算を予定し、EURTRAの開発プロジェクトが持たれているのは何故かという問に対し、「SYSTRANはアメリカで開発され、IBMのマシンを中心に稼動サービスすることに対して、やはり民族的意識から許されるものではないと考えているようである。そこで、SYSTRAN以上のものを開発することを目標にしているようだ。このこと事態、我々が批判すべきことではない。我々の仕事は、その間にSYSTRANのよりグレードアップを図ること、より良いサービスを行うことにあると考えている。そのためにも、現在日本に約25名、カルフォルニアに約20名、ヨーロッパに約50名の合計約100名の技術者を投入、システムアップおよび辞書の開発に当たっている。」とのこと。

すべての質問(調査項目)に、協力したいが時間的に無理であり、企業としての問題もあるので許してほしい、とのことで、以下、SYSTRANの概要を説明していただいた(図4-14)。

SYSTRANは、現在、IBM360, 370(OS:Mus)およびSiemens BS2000, BS3000で稼動する。IBMのUM/CMS, DOS/VSでも稼動するよう現在開発中である。SYSTRANの中心はプログラミング言語 PL/1で書いてある。

必要とするcpu容量は、仮想メモリー量を除き768KBであり、とくに、辞書の更新およびデータ保存のために2本の磁気テープ(1600/6250bpi)を要する。また3330タイプの磁気ディスク1バックを必要とする。翻訳結果および辞書出力のためのプリンタはとくに制限はない。

約100,000語の翻訳に対し、3300タイプの磁気ディスク(13.030 B)を使用の場合、磁気ディスク容量の割振りは、およそ以下の通りになる。

SYSTRAN Automatic translation procedure

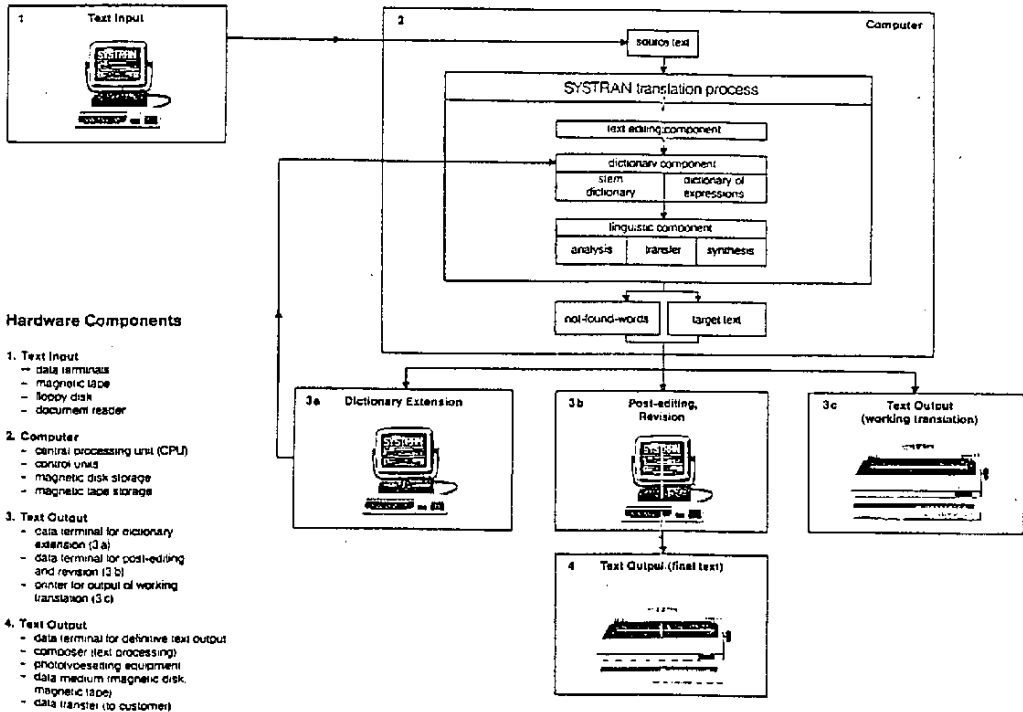


図 4-14 SYSTRANのシステム構成図

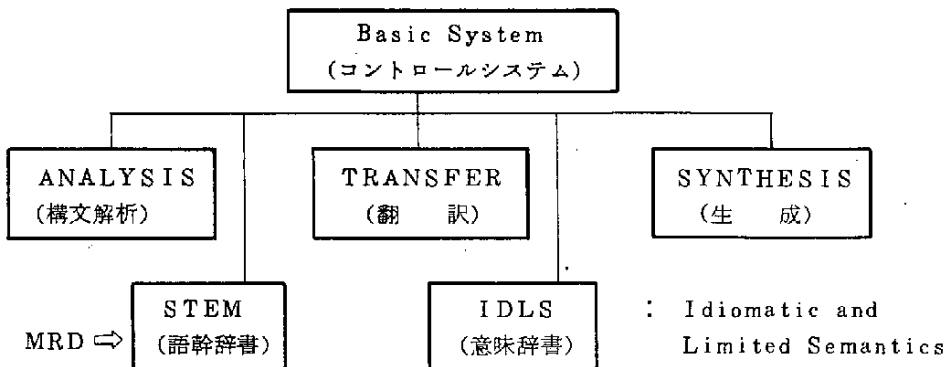


図 4-15 SYSTRANのMRDの位置づけ

- program libraries 400 tracks
- control language library 30 tracks
- libraries master version (通常はテープ) 460 tracks
- libraries, in access format (ディスクのみ) 420 tracks
- additional language pair space 80 tracks
- source and target language text 100 tracks
- work space 少くとも50 cylinders

処理時間は、pre/post editを前提として、それを含め約300,000語に対してIBM370/168使用の場合、cpu時間は1～2時間とみている。

なお、SYSTRANの持つMRDの位置づけは、図4-15のようになっている。

翻訳サービス料金は、1985年1月現在翻訳結果(35ライン/ページ、10語/ライン)をベースにヨーロッパ系言語間の対訳については以下のように設定してある。

(例1) 200ページ/月÷70,000語/月の場合

0～50,000語…… 基本料金 DM2,500.-
 50,000～70,000語(20,000語×0.045DM/語)…… DM900.-
 合計 ……………DM 3,400.-

(例2) 300ページ/月÷105,000語/月の場合

0～50,000語…… 基本料金 DM2,500.-
 50,000～105,000語(55,000語×0.045DM/語)……DM2,475.-
 合計 ……………DM 4,975.-

このサービス料金に対して、SYSTRANの月レンタル料金は、以下のように設定してある。

① BASIC SYSTEM……DM350.-/月

当システムに含まれるシステム機能

- i) Translation Control Program
- ii) Terminology Program
- iii) Text layout and Output Program

② 1対1言語間翻訳プログラム

当システムに含まれるシステム機能

- i) Analysis programs ……DM700.-/月
- ii) Transfer & SYNTHESIS program ……DM600.-/月

⑧ 辞書

i) Terminology Dictionary (約50,000語) ……DM450.-/月

ii) Dictionaries of Expression and Word-Specific Rules (約20,000語)
……DM400.-/月

以上、合計レンタル料金はDM2,500.-/月となる。

4.3.7 国際用語学情報センター

INFOTERM (International Information Center for Terminology)

住 所 : Heinestraße 38, Wien 2, Austria

電 話 : オーストリア - 0222 - 267535

日 時 : 10/21 (金) 13:30 ~ 20:00

訪問相手 : Christian Galinski (副所長)

内容:

1) まず, Terminology の必要性, 重要性, 果たす役割りについて一般的説明を受けた。

Terminology の理論は, Eugen Wüster 博士の "Internationale Sprachnormung in der Technik" (1931) の論文に端を発する。要は, 技術交流および情報交換のために, とくに専門用語を中心に用語理解を概念レベルにおいて定義する必要がある, 概念レベルで定義する方法論を INFOTERM において研究・確立してきたということ。この方法論をもとに, 各国のあるいは各機関の Terminology プロジェクトに対し, 国際的に支援 (コンサルタン ト) を行うのが, INFOTERM の重要な責務である。その基本的な枠組みは, (図 4-16) の通りである。

(資料 10) 英文

Felber, H.

"Terminology Manual" Doc, No., PGI-84/ws/21, 1984, 426p.

(UNESCO/INFOTERM)

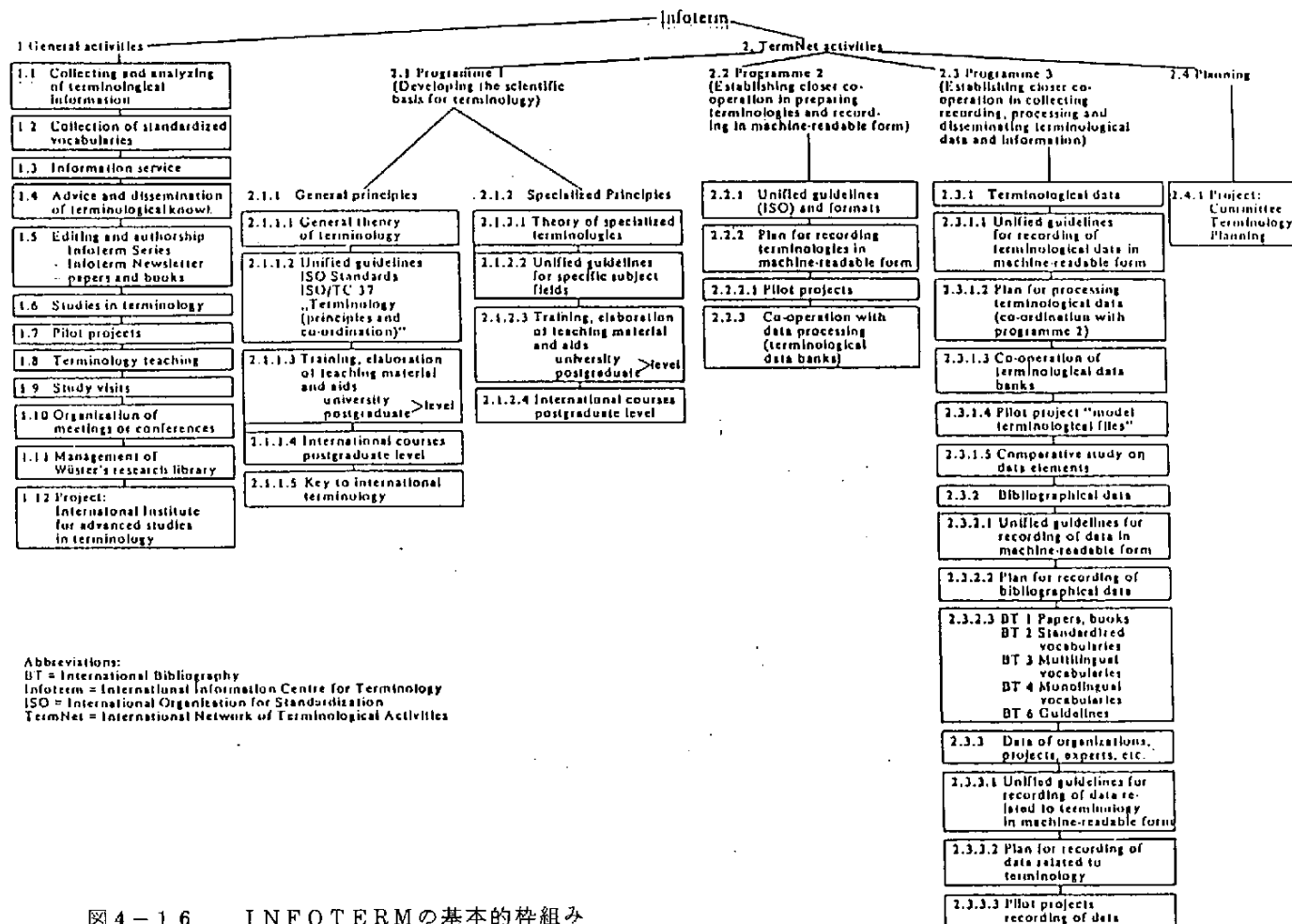


図 4-16 INFOTERMの基本的枠組み

2) Terminology databank 構想

科学技術の分野を考えたとき、1900年頭初を1としてみたとき、科学技術文献は、その後幾何級数的に発表されてきているが、その中で、新しいterminology（専門用語＝新しい知識）は現時点までに約15%増えている。（図4-16） この15%の専門用語の理解が新しい知識の理解に必要となり、terminologyに関する日常的な研究が必要になる。このように増加するterminologyを、世界的なレベルでdatabank化しておく必要性があり、現在、その基本的な枠組みについて検討に入っている（図4-17）。

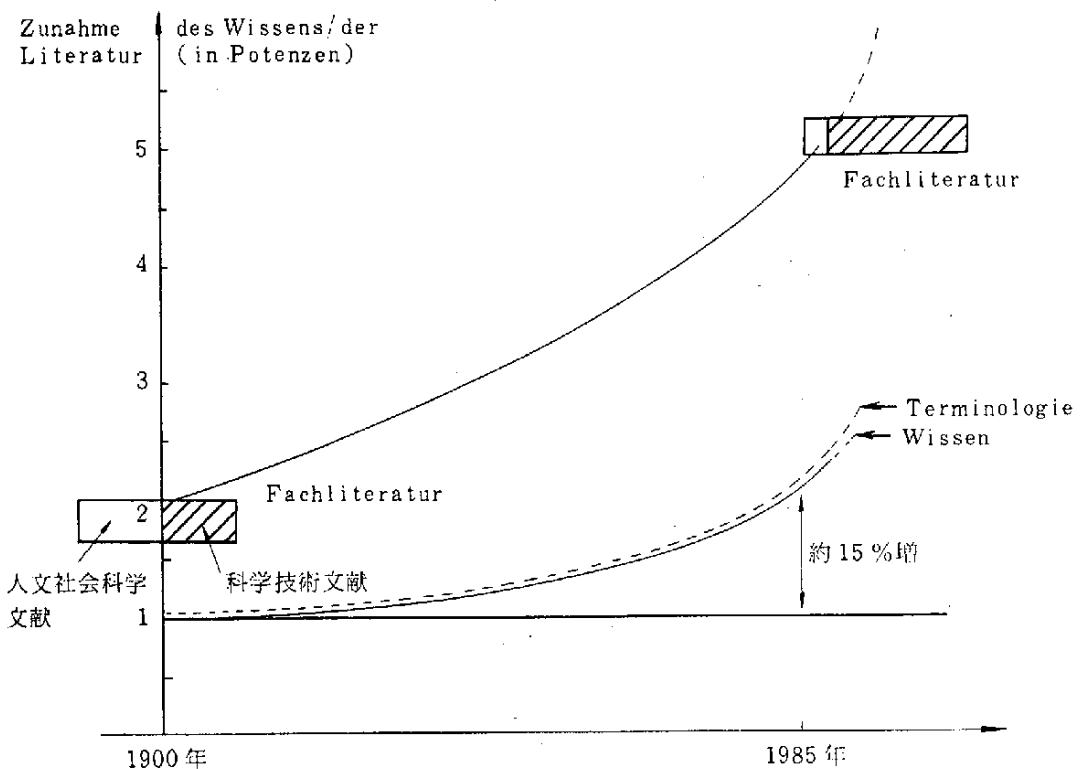


図4-16 発表文献と新しい専門用語の伸び

（資料11）独文

“Die Terminologie (databank) als Managementinstrument”

（INFOTERM内部資料）

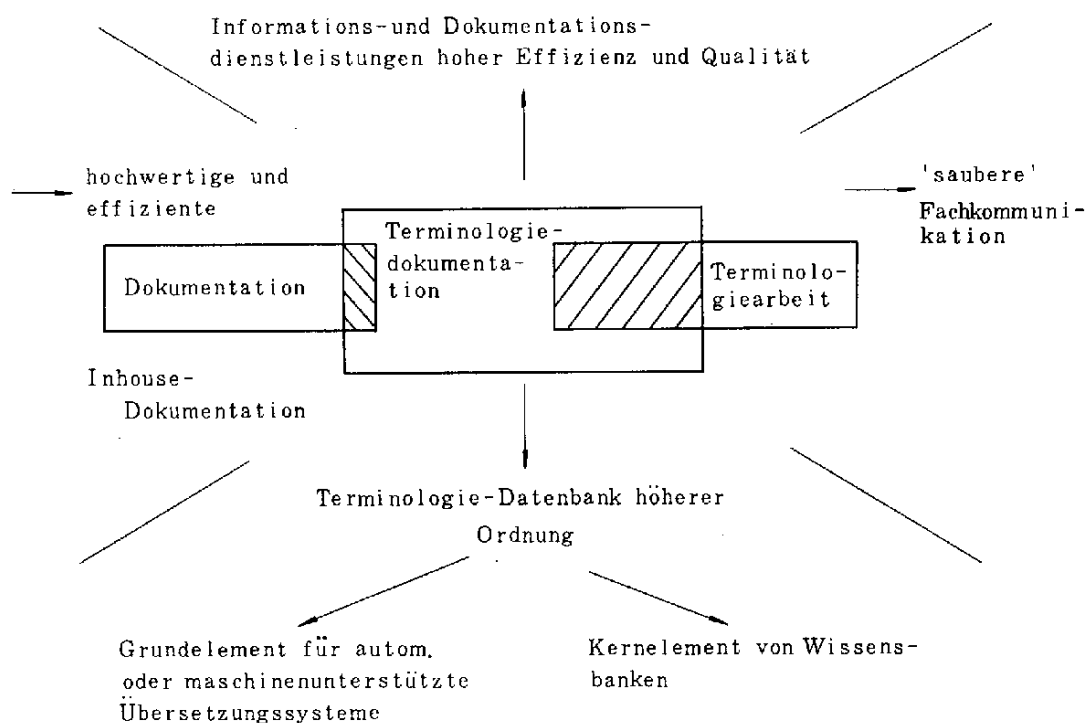


図 4-17 専門用語のデータベース化の枠組

3) MRD化の研究・開発

1984年に、EURODICAUTOMにデータを提供するために、INFOTERMのTerminologyの方法論の確立の基礎になったWüster, E. の用語集 "The Machine Tool (1968)" を入力した。図 4-18 は、その時のブルーリスト例である。この経験をもとに、現在 TerminologyのMRD化についてあるべき姿の研究を行っている。

現在、INFOTERMではECからの依頼により、電子化辞書の標準化プロジェクトを推進している。このプロジェクトは、現在ECで推進している自動翻訳プロジェクトEUROTRAと並行して、その完成においては、辞書を、各言語については各国が協力して生成・メンテナンスすることを意図して、今から、そのための標準仕様を見出そうとするものである。

(参考資料) INFOTERM

"Terminology and Related Fields 1965-1985" Hermann Böhlau
Nachf.

%%BE BTL
 %XTY WMT 68
 %XNI 0100
 %%CM N09; TE6
 %XEG
 %XVE direct reading hardness test; Rockwell hardness test
 %XDF a hardness test in which an impression is produced
 with a ball or cone indenter. The Rockwell hardness
 number is the difference E-e, e being the depth of the
 permanent indent left by the test, and E a constant
 depth used as reference. The value of E-e can be read
 directly on a dial
 %XRF Wüster: Machine Tool 1968
 %XNT UDC: 620.178.152.42; FIG: Wüster
 %XFG
 %XVE essai de dureté Rockwell; essai Rockwell
 %XDF essai de dureté dans lequel on produit une empreinte
 à l'aide d'un cône ou d'une bille. Le nombre de
 dureté Rockwell est la différence E-e, où e est la
 profondeur de l'empreinte permanente produite par
 l'essai et E une profondeur constante de référence.
 La valeur E-e se lit directement sur un cadran
 %XRF Wüster: Machine-Outil 1968
 %XNT CDU: 620.178.152.42; FIG: Wüster
 %XDG
 %XVE Härteprüfung nach Rockwell; Rockwellversuch
 %XRF Wüster: Werkzeugmaschinen 1967
 %XNT DK: 620.178.152.42; FIG: Wüster

(記号) TY: 出典
 NI: concept No.
 CM: class code
 EG: 英語
 VE: vedette = entry term(s)
 DF: definition
 RF: reference
 NT: note
 FG: 仏語
 DG: 独語

図4-18 用語集 "The Machine Tool, 1968" の入力・ブルーフリスト例

なお、MRDに関して有効な議論ができると思われる人を紹介しておきたい。

1) Prof. Jean Baudot

Dept. de Linguistique,

Univ. de Montréal

Care postal 6128, Suc. A Montréal,

Québec H3C 3J, CANADA

Tel ; (514) 343-6220

2) Dr. Thomas Schneider

Siemens AG,

Bereich Datentechnik,

Daten-und Informationssysteme

Otto-Hahn-Ring 6

D-8000 München 83

Fax ; (089) 636-2013

3) Jean M. Henning

Laboratoire d' Informatique,

Complexe Scientifique des Cégeaux

BP 45-63170 Aubière, FRANCE

Tel ; (73) 264110

4) Prof. Gest Engel

Høvedels højskole Sud.

Sønderborg, DENMARK

Tel ; (0045) 4-434225 Ex.32

4.3.8 ピサ大学

University of Pisa, Department of Linguistic

住 所 : via dello Figgiole 35, 56100 Pisa, Italy

電 話 : イタリアー050 - 502082

日 時 : 10/23 (水) 15:00 ~ 20:00

(9:30 ~ 14:30 情報学科の2人の教官の求めにより Library

automation,および Information Retrieval について情報交換を行う)

訪問相手 : Nicoletta Calzolari (言語学科 助教授)

Antonio Zampolli (言語学科 教授)

内容：

1) イタリア国立研究委員会 (the Italian National Research Council) のもとの計算機言語研究所 (the Istituto di Linguistica Computazionale) では、オンライン情報検索サービスを目的として、文献情報のデータベース (the Textual Database) の構築プロジェクトと、機械可読辞書データベース (LDB: the Lexical Database) の構築プロジェクトを推進している。今回訪問したピサ大学言語学科の Calzolari 助教授は、当プロジェクトにおいて、とくに、後者の LDB 構築プロジェクトに参加しているとのこと。

LDB の構築は、下記 15 点のようなことを支援するのに役立つと考える。

- ① 構文理解のための支援ツール
- ② データベースに対する自然言語によるアクセスのための支援ツール
- ③ 辞書構築および辞書学のための支援ツール
- ④ 綴りの修正のための支援ツール
- ⑤ 単語処理のための支援ツール
- ⑥ 辞書の出版の基礎データ・ツール
- ⑦ 自動翻訳のための支援ツール
- ⑧ 認知科学の研究の支援ツール
- ⑨ 教育支援のための支援ツール
- ⑩ 辞書研究者のための支援ツール
- ⑪ LDB の内容検索のための支援ツール
- ⑫ 文章処理のための支援ツール
- ⑬ 形態素解析のための支援ツール
- ⑭ 辞書に含まれている事項から新しい情報の抽出の研究ツール
- ⑮ 情報検索のためのシソーラスの構築のための支援ツール

当プロジェクトは、既存の冊子体辞書をまず MRD 化し、文献内に出現する単語との照合を計り、LDB を拡充している。(磁気テープ化した辞書を MRD といい、センター・マシンに登録した辞書を LDB といい、MRD と LDB を区別して使用している。)

LDB は、MSS に展開し、IBM 3081 (VSAM) で管理しているとのこと。

LDB の構成を、図 4-19 に示す。現在の容量は、LEMMAS (見出し語⇌語幹) が約 106,000 語、WORD-FORMS (語形情報⇌接辞) が約 1,000,000 および DEFINITIONS (語の定義情報⇌同義語) が約 186,000 から成り、それぞれは独立して検索できる。

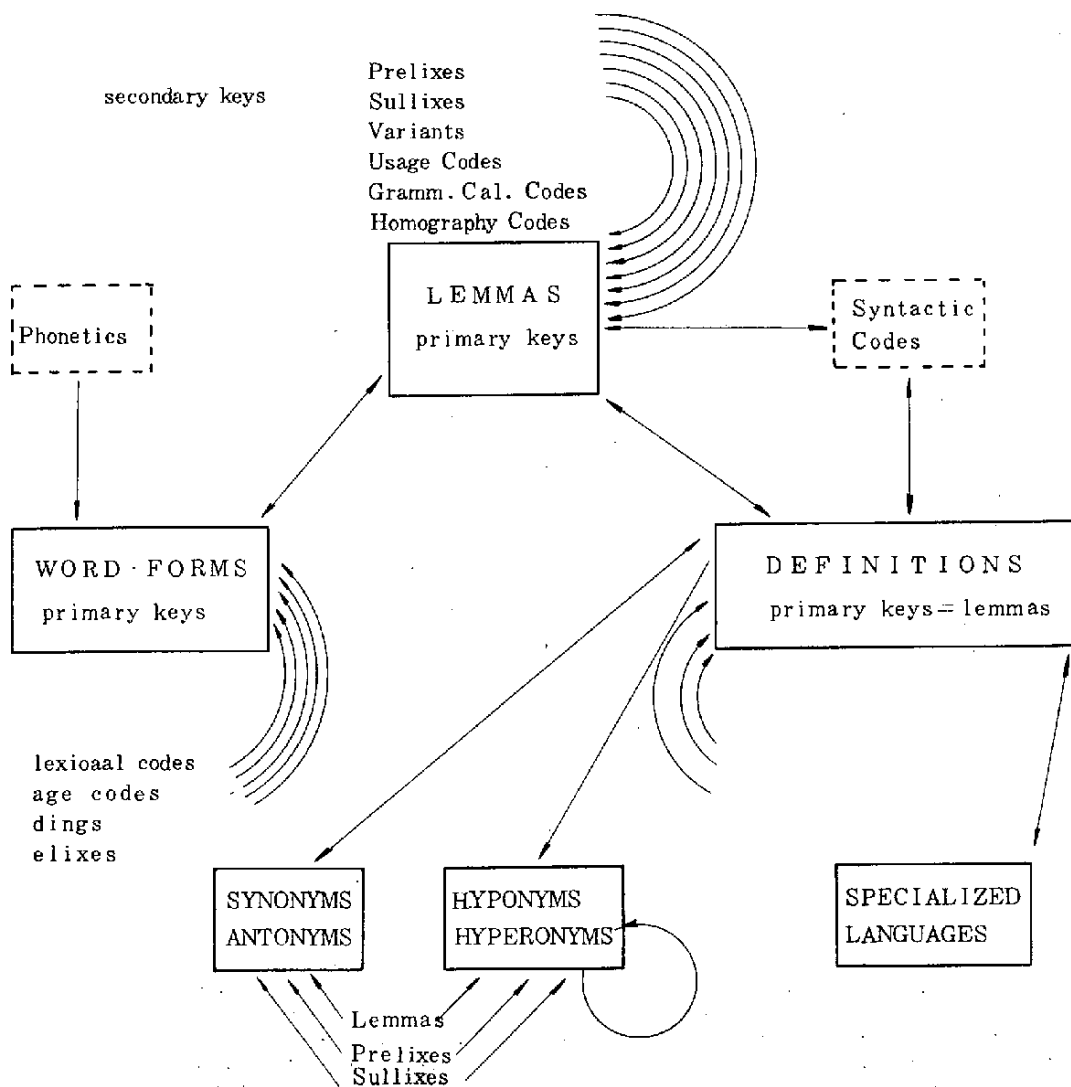


図 4-19 LDB の構成

LDBのオンライン検索は、(図 4-20, 4-21, 4-22) のような手順により利用する。

```

*****
*
* *** TEXTUAL ARCHIVE AND MACHINE DICTIONARY * N. CALZOLARI & E. PICCHI ***
* FORM : and
*      ( )      ( )      ( )      ( )      ( )      ( )
* TOTAL FREQUENCIES :      CURRENT TEXT :      Scienze e Tecnologia
* -----
* 1 - andai      P5=4; P2=2
* 2 - andaimo    P2=1;
* 3 - andando    P9=3; P5=4; P1=1;
* 4 - andante    P9=1; P9=1; P1=3;
* 5 - andare     P9=16; P5=28; P3=6; P2=13; P1=10
* 6 - andarono   P9=2; P5=9; P4=2; P3=3; P2=3; P1=1;
* 7 - andasse    P9=1; P5=5; P4=2; P3=1; P1=1;
* 8 - andassero  P5=2;
* 9 - andata     P9=6; P5=3; P3=4; P2=2; P1=3;
* 10 - andate    P9=1; P5=19; P3=1; P2=1;
* 11 - andati    P5=7; P3=2; P2=1; P1=2
* 12 - andato    P9=2; P5=9; P3=2; P2=3; P1=6;
* 13 - andava    P9=7; P5=34; P4=3; P3=1; P2=3; P1=6
* 14 - andavamo  P9=1;
* 15 - andavano  P9=1; P5=25; P4=1; P3=2; P1=9
* 16 - andera -  P5=5; P4=1; P2=1; P1=1
* COMMAND:
*****

```

図 4-20 “andare” の単語検索

```

*****
*
* *** TEXTUAL ARCHIVE AND MACHINE DICTIONARY * N. CALZOLARI & E. PICCHI ***
* FORM : andare
*      16 (P9)      28 (P5)      6 (P4)      13 (P3)      10 (P2)      (P1)
*      0.671      0.796      0.596      1.347      0.537
* TOTAL FREQUENCIES : 73      CURRENT TEXT : P9      Scienze e Tecnologia
* -----
* 1) non soffriremmo una crudele operazione senza pericolo di andare piuttosto in
*    deperimento, ed anche di mori      LV.91.rt.00.00.38..021=.004.021
* 2) credono idonei. Non si e contento di andare inestati sulla bellezza di un simile
*    risultato      AC.91.rv.00.00.40..008=A008.008
* 3) , spignendole a destra o a sinistra, fanno andare avanti o indietro le ruote della
*    vettura, secondo      AA.93.rt.00.07.26..010=0059.010
* 4) i cereali del primo arrivano alla strada maestra per andare al mercato aggravati
*    di una spesa di condotta      AC.94.rv.00.00.33..023=.A01.023
* 5) al successo dell' intrapresa. Porzioni del viaggio nell' andare e nel venire furono
*    compite in ragione di dod      AT.94.rt.00.00.34..011=0104.011
* 6) per i commerci e per le comunicazioni, da Milano andare in meno di due ore a Como
*    ed a      AT.04.rt.00.00.34..007=0106.007
* 7) total commercio rovine, e forza che a lungo andare voi, o i vostri figli, o i
*      AC.94.rv.00.00.36..034=A008.034
* 8) queste calli o viette, che spesso bisogna andare con passo da lumaca a Rialto,
*    alla Fenice      RE.94.rl.00.01.43..008=0130.008
* COMMAND:
* "X" Quit
*****

```

図 4-21 (図 4-20) に対応する短縮語義

```

*****
*
* *** TEXTUAL ARCHIVE AND MACHINE DICTIONARY * N. CALZOLARI & E. PICCHI ***
*
* FORM : andare
*
* 16 (P9)      28 (P5)      6 (P3)      13 (P2)      10 (P1)      ( )
*
* 0.671        0.796        0.596        1.345        0.537
*
* TOTAL FREQUENCIES : 73 CURRENT TEXT : P9 Scienze e Tecnologia
* -----
* LV.91.ri.00.00.38.E.
* Trovasi negli Annali della R. Societa d'orticoltura di Parigi, che un certo Pill, giardiniere
* inglese, per obligare una pianta di pero, della specie chiamata Ovo di cigno, coltivata a
* pieno vento senza mai dare un frutto, venne da un suo vicino consigliato di flagellarla ten
* bene senza riguardi: il che esegui con un bastone a tutta possa de restare non solo
* ammaccara, ma ben anche lacerata la scorza della povera pianta in tal modo punita.
* Malgrado un si barbaro trattamento la pianta anziche soffrire, diede in seguito, e sono
* gia tre anni, sempre abbondante prodotto. Questo fatto sara pur vero, e quantunque per
* tale io lo ritenga, tuttavia non mi sento punto invogliato, ne consiglierai mai chicchessia
* di seguire questi" esempio, e di ricorrere a questi mezzi bizzarri per ottenere il frutto
* dalle piante, poiche son certo che le nostre inglesamente trattare come quella suddetta del
* giardino di Vadley, non soffrirebbero una si crudele operazione senza pericolo di andare
*
* COMMAND:
* <X> Quit
*
*****

```

図4-22 “andare”のフル語義

2) 1985年10月から、あらたに Italian National Project for Automatic Language Processing がスタートし、ピサ大学言語学科では、1) の成果をもとに、a National Lexical Multiply DBの構築・整備、そのためのソフトウェア機能の研究・開発について中心的役割を果たさなければならないとのこと。当プロジェクトは、現在ECで推進しているユーレカ (EUREKA: 欧州先端技術共同研究) の一つとして、その一環をなすプロジェクトである。当プロジェクトは、ECの次の情報システムの構築のためのプロジェクトであり、EUROTRA、およびそのために必要とする各国で構築するMRDの標準化プロジェクトに対し、ECにおける情報流通のためのプロトコルの策定プロジェクトとして位置づけられる。前日(22日)にZampolli教授がLuxemburgでの会合に出席、イタリア側のプロジェクトの構想について概要を話してきたとのことである。詳細については、プロジェクトの推進が始まったばかりであり、構想の段階であり、EC各国との協調もあるので公表できないとのこと。この意味において、JIPDECのプロジェクトには大変興味があるとのことであった。

(拝受資料)

1) Calzolari, N.

Lexical definitions in a computerized dictionary

Computers and Artificial Intelligence, 2 (3), 225~233,
1983

2) Picchi, E., Calzolari, N.

Textual perspectives through an automatized lexicon
Istituto di Linguistica Computazionale (CNR)

3) Calzolari, N.

On the treatment of derivatives in a lexical database
Linguistica Computazionale, 3 (Supplement), 103~113,
1983

4) Calzolari, N.

Detecting patterns in a lexical database
Proceedings COLING '84, 170~173, 1984

4.3.9 ウォータールー大学

Center for the New Oxford English Dictionary, Univ. of Waterloo

住 所 : Waterloo, Ontario, Canada N2L3G1

電 話 : カナダ 519 - 885 - 1211

日 時 : 10/25 (金) 13:00 ~ 17:00

訪問相手 : Gayle Johannesen (事務長 女性)

J. Stubbs (Univ. Oxford Press の OED 編集部からの出向)

F. W. Tompa (計算機学科 助教授)

内容:

1984年5月から、オックスフォード大学出版部のOED編集部からの依頼により、the New OED の完全MRD化(冊子体フォーマットのMRD化)を開始した(図4-23, 図4-24)。

約3カ月間の設計の後、入力を開始した。1986年末には、すべて入力を完了させる予定である。現在、A~Az, M~Mzを完了。容量は約50MBであり、すべて入って500MBと予想している。

完成後は、学内をはじめ広くオンライン利用サービスを考えている。ともかく、現在、磁気テープ化し、OED出版部に納品することが最大の仕事である。

MRD化の目的は、OEDの編集部にあり、主に、冊子体辞書の作成支援のためであり、たとえばエントリー語の管理といったことに大きな役割が果たせると考える。他に、当MRDは、将来、

OUP-UW Joint Agreement

- OUP to capture data, ensure its accuracy, integrate the OED with its Supplements, and publish the resulting dictionary.
- OUP and UW to conduct *user survey* to help identify requirements for New OED.
- UW to design database for New OED and develop software utilities for database access and update.
- UW to explore and develop applications to exploit New OED.

図 4-23 オックスフォード大学出版部とウォータールー大学との
MRD化にあたっての取交し事項

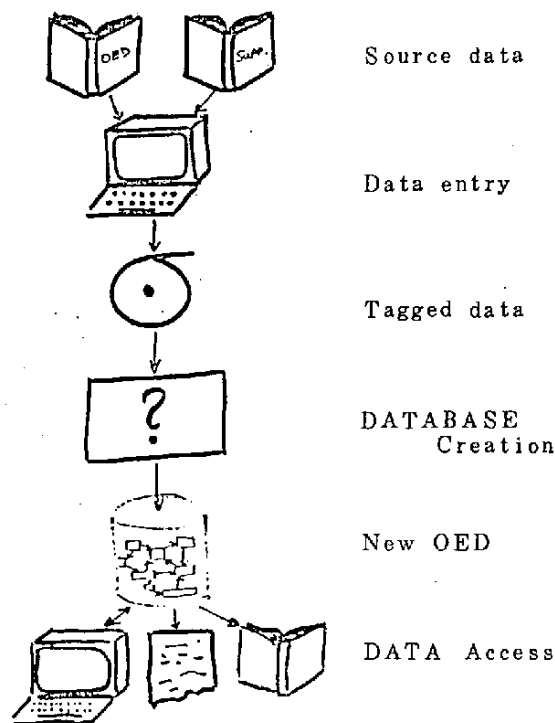


図 4-24 the New OEDのMRD化プロジェクト

文献分析、言語のコンサルティング、言語学研究に利用されていくものと考えている。これら多目的利用のために、とくに、Tompa 助教授を中心に知識ベース化の構築を考え出している所であり、たとえば、そのためのMRDフォーマットをいかにすべきかといったことについて、ピサ大学のZampolli 教授と同じくJIPDEC プロジェクトに大変興味を示している。

以下、JIPDECプロジェクトの目的、および日本でのMRD構築について説明を求められ概要を紹介した。

ウォータールー大学として当プロジェクトの受託に当り、当MRD化後の利用について具体的にどう考えているのかという問いに、プロジェクト開始に当り、図4-25のようなアンケート調査を学内を中心に行ったが、学内からの反応は少なく（というのは、まだウォータールー大学においては自然言語処理の研究は盛んでないためもある）、むしろ学外の研究者からの反応が多く、MRD化の後、利用を促進していきたいと考えている、ということであった。

（拝受資料）

Waterloo and the New Oxford English Dictionary Project.

The New Oxford English Dictionary

Oxford University Press (OUP) is about to computerize the *Oxford English Dictionary (OED)*. The entire text of the 1933 twelve-volume edition will be entered into a computer. The four volumes of the *Supplement* (the last of which will be published in 1986) will be integrated into the main text and the resulting integrated *OED* will be published at the end of 1987.

OUP plans to make the *New Oxford English Dictionary (New OED)* available in various electronic forms. The electronic version will be kept constantly up to date by revision and the addition of new entries.

The University of Waterloo (UW) is working with OUP in designing the database and creating software to access it. UW is also coordinating a user survey the objective of which is to determine the principal facilities that users will require of an electronic version of the *New OED*.

We should be grateful if you would assist us by completing the following questionnaire. Please feel free to speculate broadly on the types of facility you would wish an electronic version of the *New OED* to offer. A set of the current edition of the *OED* and the *Supplement* will be presented to five people who return this questionnaire duly completed. The five will be selected by a draw using the returned questionnaire.

Please reply, no later than 15 March 1985, using the enclosed addressed envelope.

Timothy Benbow
Oxford University Press

Professor John Stubbs
University of Waterloo



Questionnaire

1. What is your current occupation?

2. Please indicate those of your key professional activities for which the use of a dictionary is significant

- ☐ Research
☐ Writing/Editing
☐ Teaching
☐ Other (please specify):

3. Is English your first language? Yes ☐ No ☐

4. Do you regularly use a library? Yes ☐ No ☐

If so, what type of library? Mark as many boxes as apply.

- ☐ University
☐ Learned society or professional association
☐ Public
☐ Company

5. Please indicate which of the following reference works you use regularly:

	Yes	No
Oxford English Dictionary	☐	☐
Webster's Third New International Dictionary	☐	☐
A major encyclopaedia	☐	☐
Times Atlas (or other major atlas)	☐	☐
Current Contents	☐	☐
A biographical dictionary	☐	☐
Thesaurus	☐	☐
Usage Guide	☐	☐
Other (please name):	☐	☐

6. Would you regard your use of dictionaries as:

- ☐ Simple - checking on spelling, meaning, pronunciation, etc. of single words
☐ Complex - uses of a dictionary associated with, for instance, philological, literary, historical, legal or scientific studies.
☐ Both

7. How often do you use the main OED elements (listed below)?

	Daily	Weekly	Monthly	Never
Headword (e.g. for spelling)	☐	☐	☐	☐
Pronunciation	☐	☐	☐	☐
Variant forms (e.g. for historical purposes)	☐	☐	☐	☐
Etymology	☐	☐	☐	☐
Usage indicators:				
- area (e.g. Australian)	☐	☐	☐	☐
- subject (e.g. nuclear physics)	☐	☐	☐	☐
- register (e.g. colloquial)	☐	☐	☐	☐
- currency (e.g. obsolete)	☐	☐	☐	☐
- grammar (e.g. transitive)	☐	☐	☐	☐
- semantic (e.g. figurative)	☐	☐	☐	☐
Phrases and idioms	☐	☐	☐	☐
Senses (definitional)	☐	☐	☐	☐
Illustrative quotations	☐	☐	☐	☐

8. What information additional to that which is available in the current OED would you like to be present? Mark as many boxes as apply

- ☐ End-of-line hyphenation of words
☐ Pronunciation in the International Phonetic Alphabet
☐ Proper names (people, places, events, institutions)
☐ Word frequency statistics
☐ Bibliography of every source quoted (rather than the chief sources as at present)
☐ Synonyms and antonyms ☐ Illustrations
☐ Other (please specify):

9. Do you use electronic databases? Yes ☐ No ☐

If you do, are the databases you use

- ☐ Private - in-house databases compiled and maintained by your organization Yes ☐ No ☐
☐ Public - databases available to subscribers (e.g. NEXIS, LEXIS, DIALOG) Yes ☐ No ☐

10. Assuming you had ready access to an electronic version of the OED, irrespective of cost, please estimate how often you might use it (a) for simple and (b) for complex queries (see question 8)

	(a) simple queries	(b) complex queries
at least once a day	☐	☐
at least once a week	☐	☐
at least once a month	☐	☐

11. An electronic version of the OED will have a wide range of applications. The following examples have been chosen to illustrate the way in which the basic dictionary elements might be combined in order to answer specific questions.

Please indicate those which would be of use to you. Do not hesitate to mark a question which, with appropriate changes, could be adapted to the discipline in which you are particularly interested. To give a couple of examples: question 1d) could be adapted by a journalist to find examples of the suffix *-gate* used to name political scandals that resemble Watergate; question 1b) could be adapted by a chemist to produce a similar list of chemical compounds.

Obviously, information yielded by many searches of this kind would not be comprehensive but could well form a valuable point of departure for further research.

Please number the boxes according to the usefulness to you of each type of application:

very useful { 2 } moderately useful { 1 } not useful { 0 }

Broadly speaking, queries can be about (A) words themselves, (B) their sense(s), and (C) related terms:

Queries about words themselves

Historical novel-writing

1a) What interjections were in common use in the period 1670 - 1720? ☐

English language teaching

1b) List current idiomatic phrases incorporating adverbs not ending in *ly* ☐

1c) Most adjectives of the type _____ can form adverbs/nouns ending in _____. Are there any adjectives which do not appear to have such corresponding derivatives? ☐

Literary criticism

(d) List the participial adjectives that end in *-ate*, i.e. modelled on the Latin *us*. Which of them are obsolete or archaic? What is their date range? Who first used each of them?

[A: *Immolate*, obs., or arch., 1534–1830, More
Immurate, obs., 1593, R. Barnes
Impersonate, 1820–67, Keats ...]

Linguistics

(e) What is the distribution of *-lighted* as against *-lit* when the second element of a compound (e.g. *striplighted*, *-lit*)? List all headwords (main and subordinate) with either form.

(f) How many, and what kind of, words were borrowed in each of the periods specified from _____ (language specified)?

History of science

(g) List all mineral names with the dates when they were named.

Medicine

(h) List all medical syndromes, signs, etc. with the date at which they were named and the publication where the first description appears.

Botany/Dialect studies

(i) List the vernacular English names for the plants of the order *Compositae*.

History

(j) In an old manuscript document there are several illegible words. Each could be one of a number of possible combinations of letters and each combination could be an obsolete variant spelling of a standard word. Try all possibilities against *OED*'s variant lists.

Phonetics

(k) What words end with (contain) the sequence of sounds /t:stɪ/?

[A: defeatist, ..., elitist, ..., neatest, ..., retreatist, ...]

(l) What acronyms are treated like normal words when pronounced – rather than as a series of letters (i.e. like *UNESCO* as opposed to *BBC*)?

Senses of words

Literary criticism

(m) What meanings of words does Milton follow Spenser in using?

(n) What metaphorical senses of words, as evidenced by labels such as *transf.*, *fig.*, *concr.*, etc., does _____ [a writer] use, from evidence in the *OED*?

Law/History

(o) A researcher into penal history needs to know what senses of words are given definitions containing the words: 'Instrument of punishment/of torture/of execution'.

[A: Halifax gibbet, ..., joughs, ..., maiden, ..., Scavenger's daughter ...]

Law/Pharmacology

(p) What is the field of meaning covered by the word *drug* [the main entry does not cover enough ground]:

e.g. (1) What senses of words are defined using the word *drug*?
(2) What quotations are there later than 1884 containing *drug*, *narcotic*, *intoxicant*, etc.?

Law

(q) What meanings of words in the dictionary are illustrated by a quotation from an Act of Parliament?

Chemistry

(r) What terms have definitions that contain the character string ... *polymer* ... and what chemical structure does each definition give?

Related terms

History/Politics

(s) Are there quotations of the period _____ that mention the place/institution/person _____?
[E.g. Q: 1500–1600; Mozambique? A: 1588 s.v. *Inhabitant* 1.]

(t) Build up a profile of English knowledge of _____ [a place] over the period 1500–1800 from terms relating to that place.

Specialist/Technical subjects

(u) How do the coiners of terms in [a specified subject] define their terms? Display the earliest quotation for each term in this subject, where the etymology does not give a reference to a coinage elsewhere than in the entry.

(v) In what kinds of periodical are terms from statistics [or another specified subject] used?

[A: chi-square: *British Birds* ... population? [2 d]: *Biometrika*, *Jnt. Gen. Psychol.*, *Accountants Weekly* ...]

Journalism/Historical novel-writing

(w) List any quotations that associate peerage titles [Lord, Lady, Earl, etc.] with either scandal (murder, conviction, adultery, ...) or philanthropy (benefaction, bequest, orphanage, ...) so that it is possible to follow up the stories in the sources cited.

12. The applications listed in question 11 above are merely *illustrative* of the potential range of application. What other applications that would be of use to you can you envisage for an electronic *OED*? (Your answers may be as general or specific as you like.) Please answer overleaf.

4.3.10 まとめ

今回訪問調査を行った9機関について、GIDを除き、MRD構築プロジェクトの目的を、以下の2つに類別できるのではないかと認識した。

1) 自動翻訳用MRD

ザールランド大学, シュトゥットガルト大学, EC翻訳センター

2) 広い意味での自然言語研究支援ツール

ボン大学, 西ドイツ言語研究所, INFOTERM, ピサ大学, ウォータルー大学

自動翻訳用のMRD化の研究・開発は、目的語の同定について、たとえば日英翻訳のようにまったく語形の異なりが無いわけではないだけに、語根の対関係の明確化と、語尾を中心に研究がなされている印象を受けた。それに対して、とくに、ピサ大学、およびウォータルー大学のプロジェクトでは、MRDから、いかに知識データベース化を図るか、という研究が開始されたところであるように見受けられた。

そこで、今回の訪問調査の経験から、我田引水的ではあるが、わが国において、MRDの研究について、以下のような感想を抱いた次第である。

1) 自動翻訳用MRDについては、国レベルのMuプロジェクト、あるいは企業等のプロジェクトで研究が推進されている。

2) 西ドイツ言語研究所のような国語研究のためには、国立国語研究所が研究を推進している。

3) INFOTERMのようなTerminologyについては、JIS用語あるいは、シソーラスの構築の中で推進されていると考える。

以上のように考えると、残るテーマとして次のようなことがあると考えた。

1) EURODICAUTOMのように、一つのプロジェクトでは人的にも、資金的にも、時間的にも成し得ないMRDを国のレベルで開発していくことが大切である。

2) ピサ大学、およびウォータルー大学の構想のように、冊子体レベルのMRDから、知識データベース化を図る研究をする必要がある。

とくに、前者の1)のレベルについては、国レベルのプロジェクトとして構築を推進することが大切であると思えた。後者の2)のレベルについては、一つのプロジェクトで研究をスタートさせる必要があると思えた。

4.4 国内調査の概要

国内調査は、アンケート調査とヒアリング調査をおこなった。

(1) アンケート調査

自然言語解析用辞書について、昭和60年11月26日より昭和61年2月17日まで、アンケート調査回収状況一覧表(表4-3)にある24機関26辞書についてアンケート調査をおこなった。4.1に書かれている調査項目に基づいてアンケート調査(調査票は資料4)をおこなった。アンケートの内容には、先端的色彩が強く研究開発上の機密に属する事項が多く含まれており、回答に際して、企業戦略上、研究開発上、公表可能な範囲で協力をお願いした。

19の辞書についてアンケートの回答を寄せていただいた。この結果は、「4.5.3 国内調査結果の関連図表」の表Ⅰから表Ⅴにまとめられている。アンケート回答辞書の呼称は、アンケート調査回収状況一覧表(表4-3)にある辞書の略称で呼ぶこととする。

なお、残りの7つの辞書については、回答をいただくことができなかったが、それは、つぎのような理由による。

- ・「まだ、辞書が構想段階で、完成していないので、回答することができない。」(ISO, 学術用語)
- ・「機械翻訳システムとは違ってアンケートに該当することがない。」(JICST)
- ・「機密に属することが多く、書きづらい。また、現在、社内で方向を検討中であるため、今回のアンケートは、辞退したい。」(三菱)
- ・「担当者が不在のため、アンケートには回答できない。」(シャープ, 学研)

また、アンケートに対して、次のような意見が寄せられた。

- ・「当社は、出版社ですので、効率的に各種辞典類を作成するため、いわゆる電子化を進めておりますので、それらの作業過程で各種データベースが完成したという結果になっています。しかし、最近の機械翻訳システム開発の動きに合わせて、これらのデータベースをシステムの有力なツールにするため各種手直しをしております。本アンケート調査には、具体的な回答ができておりませんが、より高度のものが完成しましたら、詳細をご報告させていただきます。」

(インタープレス)

- ・「キーワード一覧は、自然語を多少整理したもので、機械翻訳に使用するような辞書とは違い、このアンケートの様式通りには、うまく回答できませんのでよりしくお願いします。」

(JAPIO)

- ・「顧客との関係で詳しくお答えできません。」(日立)
- ・「記入覧にーと記入してあるところは、記入すべき事項がないか、または、公開できないことを示します。」(富士通)

表 4-3 アンケートの回収状況一覧

分類	機械可読辞書	機械可読辞書の略称	調査対象機関	アンケート回収状況
一般辞書	(1)新明解国語辞典(三省堂)	新明解	工業技術院電子技術総合研究所	回収
	(2)学研国語辞典(学研)	学研	(株)日本電子工業振興協会	回収できず
	(3)ニューコンサイス英和辞典	コンサイス	京都大学	回収
	(4)Longman	LDOCE	京都大学	回収
専門辞書	(5)I P 電子辞典シリーズ	インタープレス	(株)インタープレス社	回収
	(6)I S O 用語ファイル	I S O	(財)日本規格協会	回収できず
	(7)学術用語ファイル	学術用語	文部省	回収できず
	(8)電字林	電字林	(株)東洋情報システム	回収
目的辞書	(9)Mu 機械翻訳プロジェクト	Mu	科学技術庁	回収
	(10)I C O T 電子辞書システム	I C O T	(財)新世代コンピュータ技術開発機構(I C O T)	回収
	(11)日本語基本動詞辞書	I P A	情報処理振興事業協会(I P A)	回収
	(12)J I C S T シソーラス	J I C S T	日本科学技術情報センター	回収できず
	(13)情報検索用辞書	J I C S T	日本科学技術情報センター	回収できず
	(14)J A P I O シソーラス	J A P I O	(財)特許情報機構(J A P I O)	回収
(商用/実験)機械翻訳システムの辞書	(15)P I V O T / V E N U S	日電	日本電気(株)	回収
	(16)A T L A S - I (英日) A T L A S - I I (日英)	富士通	富士通(株)	回収
	(17)T A U R A S (英日)	東芝	(株)東芝総合研究所	回収
	(18)メディアムバック・マイク ロバック	ブラビス	(株)ブラビス・インターナショナル	回収
	(19)機械翻訳システム	シャープ	シャープ	回収できず
	(20)A T H E N E / E A T H E N E / J	日立	(株)日立製作所システム開発研究所	回収
	(21)日英機械翻訳システム	三菱	三菱電機(株)	回収できず
	(22)機械翻訳システム S H A L T	日本 I B M	日本 I B M (株)	回収
	(23)S Y S T R A N	シストラン	シストラン(株)	回収
	(24)V A L A N T I N E , K A T E	K D D	国際電信電話(株)	回収
	(25)L U T E	N T T	日本電信電話(株)	回収
	(26)機械翻訳システム	沖電気	沖電気工業(株)	回収

- ・「ブランクのままになっている箇所は、社外秘に属するためにお答えできない部分でございますので、そのへん、御理解下さいますようお願い申し上げます。」（シストラン）
- ・「回答は、1）電字林の辞書DB、2）電字林の基本語辞書マスター、3）電字林の専門用語辞書マスターの三種類です。」（東洋情報）

(2) ヒアリング調査

アンケート調査の結果を補うべくヒアリング調査をおこなった。ヒアリング調査では、「辞書の共用使用」について自由に意見を述べていただき、アンケート調査では得られない貴重など意見を拝聴することができた。訪問先は、シストラン株式会社、国際電信電話株式会社、株式会社東洋情報システム、九州大学（吉田将教授）、福岡大学（首藤公昭教授）である。この結果については、訪問先の方の許可をいただき、掲載させていただいた（4.5.2）。

4.5 国内調査の結果

4.5.1 アンケート調査結果

アンケート調査票（資料4）により実施した、アンケート調査の結果をまとめる。具体的なアンケート回答内容は、4.5.3に図表としてまとめられているので参照されたい。

まず、4.5.1.1で、機械可読辞書がどのようなプロジェクトで開発されて、どのようなシステムで使用されているかについてまとめ、次に4.5.1.2で、調査対象の機械可読辞書が、どのような辞書であるか（どのような規模で、どのような言語情報を含み、どのような構造をもつものであるのか）についてのべる。ただし、辞書の内容に関する情報は、企業秘密に属するため、必ずしも十分な情報が得られたとはいえない。4.5.1.3では、これらの機械可読辞書が、どのようにして作成され、どのようなメンテナンスがおこなわれているかについて報告する。4.5.1.4では、エンドユーザが機械可読辞書を実際に運用するときの、問題点をあげる。エンドユーザが自ら登録した語に関する権利関係については、ひとく考え方が分かれており、社会的なコンセンサスが得られていない。4.5.1.5では、機械可読辞書の共用使用を前提とする場合の問題点、および、可能性についてまとめる。

4.5.1.1 プロジェクトおよび辞書システムの概要

機械可読（machine readable）辞書とは、計算機上に置かれている（インプリメントされた）辞書のことをいう。機械可読辞書には、人間が参照するための辞書とシステムの一部として組み込まれている辞書がある。前者の例には、新明解国語辞典（新明解）、ニューコンサイス英和辞典（コンサイス）、Longman（LDOCE）等がある。これらの辞書は、一般の冊子体辞書を機械可読にしたものである。この他に、言語研究を目的として日本語基本動詞辞書（IPA）が作られたり、辞書を引くかわりに、自動的に辞書検索をおこなってくれる辞書（電字林）が作られている。後者の例には、データベースの検索用の辞書や機械翻訳用の辞書がある。データベースの検索時に

は、キーワードとして言葉を使う。このキーワードを集めたのが、J A P I O シソーラス (J A P I O) である。また、近年、研究、および実用化が盛んにすすめられている機械翻訳システムにおいても、機械が自然言語を処理するときに辞書が使われる。科学技術庁 M u プロジェクト (M u) , 新世代コンピュータ技術開発機構 (I C O T) , 日本電気 (日電) , 富士通, 東芝, プラビスインターナショナル (プラビス) , 日立製作所 (日立) , シストラン, 日本 I B M , 国際電信電話 (K D D) , 日本電信電話 (N T T) , 沖電気工業 (沖電気) などで、機械翻訳用辞書を開発している。

機械可読辞書の開発状況は、目的の違いから多岐にわたっている。たとえば、機械可読辞書が開発されている計算機環境を例にとると、使用計算機においても大型計算機から、ミニコンやマイコンに至るまでである。さらに、実験用の辞書システムでは、特定の計算機言語のみを処理する専用マシン (Lisp マシンや Prolog マシン) も使用されている。

システムの開発用言語では、アセンブラ, C, F O R T R A N, P L / I, L I S P, E S P (Prolog + Object 指向) などが使われている。この他に、自然言語を解析するときに使う文法登録を容易にするために、独自に文法記述用言語を開発しているところもある。(シストラン)

システムの規模は、データの量で冊子体辞書で 2 0 M B ~ 1 5 2 M B, 機械翻訳用辞書では、1.5 M B y t e ~ 3 0 M B となっている。ただし、辞書の規模については、公開していない企業が多い。また、データベース検索用の J A P I O のシソーラスではデータの量が 3 1 G B となっており、このシステムは、辞書というよりも、特許情報に関するデータベースシステムといえる。

4.5.1.2 辞書の概略およびその内容

機械可読辞書開発目的別に分類すると、つぎのように分かれる。

- 既存の冊子体辞書の機械可読化 (新明解, コンサイス, L D O C E)
- 日本語研究用辞書 (I P A)
- 商用辞書 (電字林, インタープレス)
- 検索用語集, 蓄積用 (J A P I O)
- 機械翻訳システムの解析用辞書, 変換用辞書, 生成用辞書

辞書の機械可読化は、1 9 7 8 年 (昭和 5 3 年) 頃から着手されはじめ、ここ 2, 3 年のうちに開発が活発化してきた。現在までに、版がすでに何度か改定されている辞書もある。

アンケート回答の範囲に限っていえば、機械可読化されている言語は、日本語、英語、和英、英和である。英語以外の外国語の機械可読辞書は、まだ、開発されていない。東洋情報が、中国語を含めた辞書の開発を計画しているのみである。

辞書の規模は、実験や研究の目的でつくられた小規模のものから、実用化をめざした大規模なものまで開発されつつある。規模の大きい辞書では、一般用語辞書と専門用語辞書とに分けて開発している。一般用語辞書は、現在開発されているもので、見出し語で 5 万語から 1 0 万語入っており、専門用語辞書では、2 5 万語から 3 3 万語が収録されている。専門用語辞書はかなり大規模になる

見込みで、科学技術分野を例にとると、分野を一つとっても分野だけで300あり、電気の分野で2万語、学術用語で5万語収録されている。

辞書の内容は、その辞書が何の目的でつくられているかに、大きく依存し、辞書を比較する要素にもなる。つまり、辞書の中に、どのような言語的な情報が、入っているのかを示すのが、「データ項目」と「その内容」である。データ項目一つとっても、単純には、比較することができない。それは、同じ言語的な情報を、辞書によって様々な呼び方をしているからである。

各データ項目の中に、どのような言語情報を入れているかを示すものが、データ項目の内容である。この項目は、とくに企業秘密に触れる部分なので、回答を拒否する企業が多かった。Muプロジェクトの方から、詳しい資料を送っていただいた。この資料の中から、品詞のカテゴリー分類および、格マーカ、意味マーカを表に取り出してある。(Fig II-2-8~II-2-12参照)

これらのデータ項目を、計算機の中にどのような形式でいれておくかを決めるのが、データフォーマットとファイルフォーマットである。データフォーマットをどのような形式で結び、辞書にどのような構造をもたせるのかをきめるのが、ファイルフォーマットである。ファイルフォーマットは、辞書引きの速度やメモリー効率に影響が出るとともに、辞書引きの目的に応じて決められる。三省堂の辞書では、固定長のランダムファイルが使用されており、IPAの日本語基本動詞辞書では、順編成ファイル、機械翻訳システム(東芝、日本IBM)では、可変長のファイル構造を採用している。ファイルについて、東芝では、「必要に応じて変換可能」と述べている。

4.5.1.3 辞書の作成およびメンテナンス

(1) 機械可読辞書の開発方法

機械可読辞書の開発方法には三つのタイプがある。

① 冊子体辞書を機械可読辞書にしたもの

このタイプの辞書の開発には、入力、メンテナンス作業のみを行えばすむ。言語情報を収集したり、分析したりする作業は行わずにすむ。今迄に、このタイプの辞書から、機械翻訳システムのための辞書作成の研究が行われている。

② 当初からそのシステムを動かす目的で開発された辞書

システム開発を「主」目的とする場合には、辞書の開発コストをできるだけ低く押さえないし、また、ユーザにすぐに使わせるためには、マンマシン・システムを備えて、簡単に辞書のメンテナンス作業が行える必要がある。機械翻訳システムのために開発された辞書の多くは、この部類に属する。ユーザは利用しやすいが、汎用の機械可読辞書にはなりにくい。

③ 汎用に機械可読辞書をめざして開発されている辞書

システムの辞書として開発されているが、まず、汎用型の辞書をつくり、辞書管理システムにより、システム用の辞書がつくられる。このように、辞書管理システムがしっかりしているものは、汎用機械可読辞書になりえ、応用が広い。

辞書作成には、まず、情報処理技術者と言語学・国語学の専門家が辞書の仕様を決定し、その仕様に基づいてプログラマ、インプットオペレータが作業を担当しているケースが多い。規模の大きいものについては、辞書のデータ作成・編集のために10名～数10名の人員を要しており、入力作業などについては、外部に依頼することが多い。

(2) 辞書の作成手順

辞書を作成するには、各調査機関ごとに、どこに力点を置くかの差はあれ、次のような作業手順を踏んでいる。

- ① 言語データの収集
- ② 言語データの分類・分析
- ③ 言語データの入力
- ④ システムの利用しやすいフォーマットへの変換
- ⑤ 修正

単語の収集についていえば、既存の辞書の場合（新明解、コンサイス、LDOCE）は、とくに作業する必要はないが、国立国語研究所の調査結果および種々の市販辞書等から収集をおこなっている。機械翻訳システム用の辞書では、実データに基づいて語の補充（東芝）がおこなわれている。また、収集作業のほかに、システムを効率的に動かすために見出し語を調整（富士通）する必要もある。

収録単語に言語的な情報をどのような尺度で分析し、付加するかということが問題になるのだが、その方法論が厳密に確立されていないためか、収録単語の分類作業については、あまり回答が寄せられなかった。

収録する単語の収集から分類、各種のマーカの付与に至る作業手順について、つぎのようなりかたがある。

- ① 辞書登録の際に、インプットオペレータと画面上で対話的におこなうことにより、各種マーカを付与する。（ブラビス）
- ② 単語の収集、各種マーカ付与にKWIC（同音語辞典）を用いる。名詞に対する各種マーカの付与は、辞書や用語集を参考にし、用言のマーカ付与については、KWICなど出現した文を中心に分析を行う。言語的知識をもつ人が、詳細な作業マニュアルに基づいておこなう。

（Mu）

また、収録単語の選定基準として、「頻度」を挙げるところが多い。その頻度の決め方には、市販の辞典に収録されている見出し語をベースにして、調整をはかっている場合もある。

入力作業については、Muプロジェクトのように詳細な「データエントリー・ルール」を決めておこなっているものもあるし、会話形式でおこなうため、データエントリー・ルールが不要なシステムもある。Muプロジェクトのデータエントリー・ルールを参考資料として、表にあげて

おいた (Fig III-3-6)。

(3) メンテナンス

辞書は、常にメンテナンスを必要とする。その方法は、以下のようにしておこなわれている。

冊子体辞書では、インプットミスの修正を定期的に継続的におこなっている。冊子体辞書の原本の版が変わるごとに、機械可読辞書の改定もおこなわれている。

システムの一部として組み込まれている機械翻訳用辞書にとっては、メンテナンスは頻繁になる。機械翻訳用辞書には、ユーザが専用に使う私用 (ユーザ) 辞書と、原本の辞書、すなわち、メーカーの辞書がある。メンテナンス内容も私用 (ユーザ) 辞書についておこなわれるものと、原本の辞書、メーカーの辞書についておこなわれるものがある。ユーザの私用 (ユーザ) 辞書についておこなわれるメンテナンス内容には、新しい単語の登録やインプットミスの修正などがあり、ユーザにより随時おこなわれる。また、原本の辞書、メーカーの辞書についておこなわれるメンテナンス内容には、インプットミスの修正、辞書内容の更新、対訳の適切チェックなどがあり、これらは、定期的 (3 カ月、6 カ月～1 年、版改訂のたび) におこなわれている。メンテナンスは、ユーザ辞書を編集・修正し、その結果がメーカーにいき、ユーザではやりにくい言語的知識やシステムに関する知識を要する作業 (①追加登録すべき語の洗い出し、②誤りの洗い出し、③入力、④チェック) を行い、修正されて、ユーザへ修正をフィードバックするという手順でなされている。

機械可読辞書作成、メンテナンスにおける問題点や意見として、次のようなものが寄せられている。

① データの管理体制

- ・ データが途中で人手に渡り、汚いまま拡散して しまった。きちんとしたマニュアルの整備が必要である。
- ・ 専門委員会により一元管理する体制をとって、効率的な運用をはかっている。

② マンパワーの問題

- ・ 入力および入力された語のチェックをするのに人手がかかるので、これらの簡易化をはかる必要がある。

③ 基本語辞書の問題

- ・ 基本語辞書といっても、完全に分野独立の用語は少なく、ユーザによっては、相反する要求がでてくる。これをどのように調整するかが課題である。

④ 冊子体辞書と翻訳用辞書の問題

- ・ 冊子体辞書は、必ずしも翻訳用にはできていない。冊子体辞書から翻訳用辞書をつくるには、どのようにして情報を付加するかが問題である。

4.5.1.4 辞書システムの運用

(1) 運用上の問題

辞書システムは、電気、情報、医学、原子力、自動車、鉄鋼、金融などで使用されており、実際にこれらのエンドユーザが、辞書システムを使い始めると、運用上の問題がおこる。例えば、辞書に載っていない用語や、ユーザ固有の言いまわしなどができて、未登録語の登録および、辞書内容の修正という作業が、必要になる。この作業を支援するために、システムに辞書エディタや辞書ユーティリティを整備したものや、対話的に更新、修正ができるように工夫してあるものがある。辞書に対して、エンドユーザが手を加えられるのは、ユーザ辞書のみで、基本語辞書は、変更させないものが多い。

また、辞書内容の変更には、言語的知識やシステムに関する知識が必要となるので、エンドユーザで起こった問題を集めて、辞書を修正してエンドユーザに返すというフィードバック体制を敷いているところもある。(Mu, シストラン, 富士通)

(2) エンドユーザの権利

エンドユーザが追加・修正した辞書に関する権利関係は、現在のところ考え方が統一されておらず、混乱のもとになりやすい。エンドユーザにより追加・修正された部分の権利関係については、以下のように意見が分かれている。

① 契約に基づく、あるいは協議をする。

- ・ 辞書提供時の著作権者、版權者との契約に従う。(新明解)
- ・ 排他的権利を設定しようとする場合、事前にIPAと協議する。(IPA)

② エンドユーザに帰属する。(ブラビス, ATLAS, 電字林)

③ エンドユーザの権利を認めない。

- ・ 原則としてユーザの占有権は認めず、シストランが権利を所有する。権利の一本化をはかり、第三者に利用できる形態にしてある。ただし、ユーザの企業秘密に属する場合は例外である。(シストラン)

④ エンドユーザとメーカーで折半する。(日電)

⑤ 未検討

- ・ 一元管理用辞書は、今後設置される委員会等の承認を伴った上でおこなわれる。個人用辞書(ユーザ用辞書)については、まだ、考えていない。(Mu)

4.5.1.5 辞書の共用使用

アンケート回答者にとって共用使用のイメージが、漠然としている点が問題である。冊子体辞書から機械翻訳用辞書をつくることを考えている人が多いようである。また、アルゴリズムとのかかわりをもつ機械翻訳用辞書を互いに共用することの難しさを指摘する人が多い。

(1) 共用使用の可能性とその問題点

共用使用の可能性について、一般辞書と機械翻訳用辞書とで大きく立場がかわる。

つまり、情報を提供する側の意見は比較的楽観的である。例えば、コンサイスやLDOCEでは、「辞書そのものは、検索以外の目的に適していないので、機械翻訳システムなどに直接使用することは不可能であろう。しかし、色々な工夫をすれば、システムの目的にあった情報を抽出することは可能であろう。」また、IPAでは、「利用契約書による」といっている。

一方、情報を提供される側は悲観的な意見が目につく。例えば、機械翻訳用辞書をもつところでは、「共用使用はできない」、もしくは、「かなり難しい。」というところが多い。この難しさには、次のようなものがある。

① 変換の問題

現在、データフォーマット、コード、使用言語などは開発主体ごとに独自に決められているので、共用使用するためには、変換という問題が生じる。(東洋情報, Mu)

② 辞書とアルゴリズムの関係

見出し語、数、性、人称以外はシステム固有のものであり、一般に共用できない。たとえば、品詞、意味情報は、パーザ・意味分析器と一体のものである。(東芝)

③ 言語的な記述情報の体系のちがい

システムのアーキテクチャによって、辞書の記述情報の体系がちがう。その中で共通項を見出すのは非常に困難であろう。(シストラン)

④ 権利の問題

共同版權であるため。(東洋情報)

(2) 共用使用する場合の辞書の作成、および、メンテナンス上で考慮すべき点

① ユーザの組織をつくり、情報を伝達する。(新明解)

② 辞書項目に対する修正の履歴を各項目中に記述する必要がある。(Mu)

③ 本辞書に与える影響は少ないが、本辞書が改訂されれば、影響が及ぶであろう。(コンサイス, LDOCE)

④ 編集要領を統一的にしたり、フォーマット等を特殊にしておかなければならない。

⑤ logical formatの形式から、他のシステムの辞書を作成する。(ICOT)

⑥ 共用使用に対して否定的な見解(東芝, シストラン)

(3) 辞書の提供可能性とその条件

アンケートの調査対象辞書を外部に提供する可能性はあるか、また、あるとすれば、どんな条件のもとで提供するかという問いに対する回答には、以下のようなものがある。

① 無償提供または無償貸与

新明解は、著作権者や版權者と契約書の交換をする必要がある。

(ケ) 提供に際して、ユーザとIPAの間で契約書を取り交わすことになっている。

i) 辞書の複製の禁止

iii) 利用状況についてIPAに報告する。

(ウ) 提供方法は、磁気テープおよび「辞書の使用説明書」を貸与する。

磁気テープの複写代金等の実費はユーザの負担とする。「辞書の使用説明書」は、辞書貸与ユーザに対してのみ提供する。

ニューコンサイスは、三省堂が直接有償提供している。

インタープレス社は、販売している。相手先の利用目的に応じて、契約条件を設定している。

電字林はすでにパッケージ商品化している。条件については未検討。

富士通はATLASのユーザに限り、提供。

日本電気は、コピーの禁止を前提として、提供可能性を示している。提供媒体は、ROMと
のこと。

シストランは、提供可能性ありとのこと。

Muは、未検討ではあるが、特定の個人団体が利益を得ないことを条件として、低価格で提供すべきだと考えている。

ICOTは、未検討ではあるが、提供可能性ありとのこと。

ブラビス, 東芝

4.5.1.6 参考資料

アンケート調査の回答の中で、「関連資料および参考文献」として記載されていたものを以下にリストアップして参考に供する。ここで辞書名の前の番号は表4-3に準じている。

(1) 新明解国語辞典（三省堂）

- ・ 横山晶一「国語辞典磁気テープのドキュメント」電総研集報（1984）

(3) ニューコンサイス英和辞典（京大長尾研）

- ・ Nagao, M. et al : An Attempt to Computerized Dictionary Data Base, Proc. COLING 80, 1980, 9.
- ・ 長尾真 他「英和辞書データベースの作成とその応用」情報処理第21回全国大会, 1980, 5.
- ・ 長尾真 他「コンサイス英和辞典に現れる日本語単語の性質」情報処理第23回全国大会, 1981, 10.
- ・ 上田良寛「英和辞典のデータベース化に関する研究」京大・工・電・卒論, 1980, 2.
- ・ 上田良寛「新コンサイス英和辞書『利用手引書』」

(4) Longman (LDOCE)

- ・ 長尾真 他「機械翻訳に対するロングマン辞書データベースの応用」情報処理NL研 29-5, 1982. (資料3)
- ・ 谷口慎也「ロングマン英英辞典のデータベース化に関する研究」京大・工・特研報告書, 1981.
- ・ 藤田公一「ロングマン現代英英辞典の特殊コードを用いた英文解析」京大・工・特研報告書, 1982.

(8) Mu機械翻訳プロジェクト

- ・ 長尾真, 辻井潤一, 中村順一, 坂本義行, 島海剛, 佐藤雅之「科学技術庁機械翻訳プロジェクトの概要」情報処理学会誌, Vol. 26, №10 (Oct. 1985).
- ・ 辻井潤一「辞書の構成と機械翻訳」情報処理学会誌 Vol. 26, №10 (Oct. 1985).
- ・ 長尾真, 辻井潤一, 中村順一「日英科学技術文献の速報システムに関する研究概要」情報管理, Vol. 28, №1 (Apr. 1985).
- ・ 坂本義行, 有賀妙子「Muプロジェクトにおける総合システムの基本設計」情報処理学会, 自然言語処理研究会, 46-6 (Dec.).
- ・ 日本科学技術情報センター, 電子技術総合研究所, 京都大学「日英科学技術文献の速報システムに関する研究, 日英科学技術用語辞書データベースの開発に関する報告書」昭和60年3月

- ・ 電子技術総合研究所, 京都大学「日英科学技術文献の速報システムに関する研究, 言語処理システムの開発に関する報告書」昭和60年3月
 - ・ 工業技術院総務部計画課(情報計算センター)「日英科学技術文献の速報システムに関する研究, 総合システムの開発に関する報告書」昭和60年3月
- (10) ICOT
- ・ T. Ishiwata, H. Tanaka, S. Amano, H. Uchida, T. Ogino, H. Miyoshi, Y. Tanaka and T. Yokoi "Basic Specifications of the Machine-Readable Dictionary" ICOT technical Report : TR-100
- (11) 日本語基本動詞辞書IPA
- ・ 「ソフトウェア文書のための日本語処理の研究-5 — 計算機用レキシコンのために—」IPA技術センター報告書, 1983年・12月
- (16) ATLAS-I (英日), ATLAS-II (日英) 富士通
- ・ 「FACOMOS N ATLAS-I解説書」
- (17) TAURAS (英日)
- ・ 天野, 平川「英日機械翻訳用パーサについて」情報処理学会, 自然言語処理技術資料 NL32-1, 1982
 - ・ 天野, 平川「機械翻訳システム」同上, NL34-1, 1982
 - ・ 天野, 他「LEXICAL NETWORK GRAMMAR」電子通信学会, 59年度総合全国大会6-60, 1984
 - ・ 天野, 他「語彙遷移ネットワーク文法について」情報処理学会, NL49-4, 1985
- (20) ATHENE 日立
- ・ 平井, 他「英和機械翻訳システム ATHENE-82辞書保守システムについて」情報処理学会第27回(昭58年後期)全国大会 6F-6
 - ・ A. Okajima et al. "Lexicon structure for machine translation an example from English-into-Japanese translator ATHENE" Proc. of ICTP'83, 1983 (Oct.)
 - ・ Y. Nitta, A. Okajima, H. Kaji, Y. Hidano, K. Ishihara "A proper treatment of syntax and semantics in machine translation." COLLING 84. PP159-166. (1984. 7)
 - ・ 梶, 岡島「日立における機械翻訳システム」情報処理 S60/10.
 - ・ 梶, 吉村, 白井「日英機械翻訳システム ATHENE/Nにおける文法記述言語」情報処理学会, 第31回全国大会, PP. 1347-1348 (S60/9)

② IBM

- ・ 「IBM情報処理用語英和対訳集」(第二版)
- ・ 堤「計算機マニュアルを対象とした英日機械翻訳について」IBM REVIEW 89 (1983)
- ・ 堤, 他「英日機械翻訳システム・SHALTの概要と英文解析および英日トランスファー」情報処理学会 第31回全国大会
- ・ 堤, 他「英日機械翻訳システム・SHALTにおける単純名詞句の翻訳」情報処理学会 第31回全国大会

4.5.2 ヒアリング調査結果

4.5.2.1 シストラン㈱

昭和61年2月3日

マネージャー 神 力 望氏
チーフリングリスト 赤 沢 江里子氏

(1) 体系づけ

JIPDECプロジェクトの構想も、基本構想があれば答えられるのだが、にわかには、よく構想がわからないので、何とも答えようがない。各システムでは、手法・記述内容もかなりちがうので、3年間のプロジェクトで、共通項を見い出すだけでも大変ではないか。各辞書の情報を対応づけることはできても体系のちがいという問題が残るだろう。すべての情報をもつスーパー・エントリーをもつ辞書を作ると、コスト面で工数がかかりすぎる。一番大きな問題は体系の問題である。体系づけをしようすると、どのシステムにも役に立たないものができる可能性がある。

(2) 運用体制

シストラン社では、すでに15～6年の歴史があるので、システム上の問題はクリアーされている。辞書は、ルール付辞書(対応する訳語をつける)を除いて、ユーザが登録できるようになっている。シストランでは、ユーザとのフィードバック体制を整えニーズに対応している。

(3) 国に期待すること

<この言葉はこの訳す>と確定していない言葉が多いので、言葉の規格集、ターミノロジーデータベース(terminology data base)を行政に期待する。日本国にとって多言語のターミノロジーデータベースがあればさらによいだろう。

なお、国で開発した辞書の公開にあまりきびしい制限をつけない方がありがたいと思う。

第1特別研究室 武田 学氏
楠 博 史氏

(1) 公表できない理由

少人数による研究段階であり、入力方法やメンテナンスの手順などが確立されておらず、公開できる状態ではない。

(2) 共用について

テクニカル・ターム（専門用語）、固有名詞（地名・人名）等は、語数が多いし（数十万）、訳語が一对一に決まる。こういった語に関する辞書があれば、使いたいし、そういうものは、共用した方がよいだろう。一般語は加工の素材となるようなものがあれば利用したい（例：IPAの辞書、森田著『基礎日本語1・2』角川小辞典シリーズ）。将来的には国が1千万語規模の辞書を作成・管理し、各システムの辞書はそれを素材として加工・作成できるようにすることが望ましい。

(3) JIPDECプロジェクトについて

まだ変動のはげしい時期で、議論も鎮静化しておらず、企業の協力も得にくいのではないかと。

応用技術システム開発第一部 小 木 正 三氏

(1) 共用について

小規模でしっかりしたものを発展させて行く方法が安全である。各社の辞書の共通項を取る方法も考えられるが、自社の辞書については、共用を前提として公開したくない。標準辞書を最初から作成するとコストがかかり過ぎ、それが使用料に影響すれば普及を妨げることになるだろう。

(2) 作成・管理について

入力段階での転記ミス、パンチミスが多い。スペルチェッカー等の支援システムが必要である。メンテナンスは常時継続されるため、管理システムが重要になる。入力前のデータの8割近くは変更、修正がある。機械可読だと、操作が容易であるので、入力後の再分類は頻繁になった。

(3) 著作権

個々の訳語にまで著作権がおよぶのだろうか。辞書の内容についての著作権は曖昧である。

(4) JIPDECプロジェクトについて

競争状態は良い結果をもたらさないので、何らかの標準化は必要だと思う。

(1) 大学での研究体制

大学関係では、一部の大学を除いては、自然言語処理の研究をやろうとしても、辞書などの蓄積を要するので、研究がスタートしにくい状態である。

九大では辞書を公開している。この辞書には、見出し語が8万3千語あり、文法に関する情報が入っている。意味に関する情報は分類情報以外はほとんど入っていない。この辞書は、昭和40年頃からつくり始められたもので、九州芸術工科大学の稲永氏の努力によるところが大きい。

辞書の開発には、非常に大きな労力とお金がかかる。辞書を良くするために大切なことは、「自由に使わせる」ことである。辞書に限らず、ソフトウェアもそうだが、「自由に使わせる」ことが物事を良くする基本だと思う。

辞書の管理については、辞書が誰に渡っているのかは知っておきたい。ユーザは、自分の責任で自由に辞書を編集・情報加工して使うのがよい。何故なら、皆、使い方が違うのだから。辞書を最初からつくるよりは編集し直した方が早いであろう。

大学で辞書開発をやると問題が多い。辞書を作っても、学術論文にはなりにくいいため、研究者としての評価がされにくい。参考文献として、辞書の「利用の手引き」を書いてもらっている。こういった形でしか評価できていないのが現状である。

(2) 共用辞書について

現状の辞書開発状況は、バラバラであるが、共用辞書はどうあるべきか?

大源の汎用の管理システムをもつ「源辞書」をつくるべきである。それをユーザがどう加工するかは自由である。もし「源辞書」にない情報が必要になった場合は、「源辞書」にとりこむという形で「源辞書」をつくれればよい。

どこかの機関がやらなければ、できないと思うが、自由に使えて、目的に応じて加工することができることが大切である。

三省堂の機械可読な辞書は随分研究に役に立っている。使わせることによって、辞書の誤まりが見つかり、訂正・改良にも役だつので、自由に使わせて欲しい。

辞書は、ユーザが加工・編集して使うものである。記述体系のちがいは、作成者によって、どうしてもおこるものである。例えば、意味分類カテゴリーを定め、共通項の設定をしたとしても、そのワーキング・グループの考えでしかない。しかし、IPAの辞書と我々の辞書を比較してみた結果、意味コードについてかなりの部分は一致していた。共通項のみをみいだすことが、大切なのではない。多義語の分類については、今後の基礎研究が必要である。この人の分類ではこうというものを、「源辞書」にすべて入れておかざるをえない。格の設定についても、皆違うのが当たり前である。似てはいるが違う。

最近、京大の辻井氏が言っているように格の種類にしても、中間の格があり、そのまた中間の格が考えられるわけである。人によってそれが大事だったりする。連続に意味が変化しうるとい
うのが、言葉の本質というものだと思う。

(3) 協力体制

- ・ どういう協力体制が考えられるか？

IPAやJICSTのように、言語学者の方々に協力してもらうのがよいだろう。とりま
め役は、工学関係者がよいだろう。

(4) 国に望む点（言語政策）

日本の財産として辞書をつくり、自由に使わせることが大事だ。そのためには、国の事業とし
て永続的な機関が関与することが必要であろう。ユーザの独自の要求があれば、ユーザが自分で
変えていけばよい。

海外には、用語を収集する機関があるようだが、国として、補助する必要があるのではないか。

国でつくった辞書が、海外に流出される可能性はあるかもしれないが、辞書はいつでも、更新
されていなくては、使いものにならない。informatics（情報学）という英語があるが、こ
の言葉は、英語を母国語とする人達にとっては、あまり響きのよい言葉ではないらしい。どうも
英語圏外でつくられた言葉らしい。将来、海外で誤用・創作された日本語が入ってきて、それ
が幅をきかすということの方が恐ろしいと思う。国際化時代における日本の言語政策上、体制を
考える必要がある。

新しい用語や学術的な term は、いくら管理しても、どんどん出てくる。きちっと整理された
terminology data bank（辞書とは、ちょっとちがうかもしれないが）が必要である。

(5) JIPDECプロジェクトへのアドバイス

多額の開発費でなければ、実際に辞書を開発するのは難しい。データ項目について、きちんと
整理をして、こうすると提言するところまでいくにも、それなりの費用がかかろう。

現在の辞書は、見出し語に対する情報が記述されているが、見出し語と見出し語の関係は、
（implicitに含まれているが）まだexplicitには書いていない。シソーラス（上位語・下
位語）の関係もとらえにくい。共用辞書は、人間が使うための辞書の改善に役立つように考える
べきではないか。つまり、「源辞書」は人間用にも、機械用にもどちらも生成できるようにすべ
きである。

(6) 参考文献

- ① 稲永紘之「日本語処理のための機械辞書」情報処理第23巻第2号 昭和57年2月
- ② 吉田将，日高達，稲永紘之，田中武美，吉村賢治「公用データベース日本語単語辞書の使用
について」九州大学大型計算機センター広報別刷 Vol. 16 No. 4, 1983

(1) 大学での研究

私のところの辞書は、日本語処理のために、関係表現・助述表現を中心とした、分野に依存しない表現を収録したものである。表現の接続関係、活用関係のルールを併用して、形態素処理、構文・意味処理に利用できる。16, 7年前から研究を続けており、最近データのばらつきがなくなってきた。辞書の語彙数は、見出し語が約6000入っており、現在、原則として公開していない。以前に公開し、かなり商業ベースで使われているようだが、我々のところには何のフィードバックもない。辞書の研究は、マンパワーをかけても、研究業績になりにくい点に問題がある。辞書の構成論もテーマになりやすく、内容全体提示もやりやすく、辞書を使った処理システムの評価に依存してしまうので、研究面では、議論しにくい。研究のすすまないのは、その辺に原因がある。何か辞書の評価基準を確立できないものだろうか。

拡張文節モデルの発想は、意味の単位で語をとらえなおそうというものである。この考え方をを使うと、形態素処理の段階からdisambiguityをおこなえ、構文・意味処理とのつながりも自然である。たぶん、形態素処理については、一番いいデータをもっているのではないかと自負している。現在、決定版をVAX750上で実現し、来年度、その仕上げをしようと思っている。

今、欲しいものは、ランダムハウス級の大量のデータを持つ対訳辞書と「生の文」である。数万センテンスの生の文が欲しい。日本語の生の文を最低5万センテンスくらいかけてみないと、安定しないようである。学会で、標準的なコーパスを何種類かもつべきである。3千文くらいはつくったが、そんなものではだめである。

かわった表現、たとえば、「山をはる」とか比喩的な表現とその用法を調査している。まだ、完全ではないけれど、収集をおこなっている。

辞書だけあっても、updateしなくてはだめである。

私は、言語現象の方に興味があり、言語のパフォーマンスを大切にしている。現象に根ざしたアプローチをしなくてはならない。言語処理研究の難しさは、primitiveのオーダーが、従来の自然科学とは違って相当多くなる点にある。

(2) 公表したくなるような環境づくり

辞書づくりについての基準をつくるというのも難しいだろう。それよりも、皆が辞書の内容を公表したくなるような環境をつくるのが大切だろう。情報処理学会あたりでも、辞書づくりをやっている研究者の努力に報いるようにしてもよいのではないか。それも、評価基準をかなりゆるめに設定してもよいのではないか。辞書づくりの努力を認める学会をつくってしまおうという話もあるが……。

(3) 参考文献

- ① 首藤公昭「専門分野を対象とした日英機械翻訳について」情報処理第14巻第9号
昭和48年9月15日
- ② 首藤公昭, 鶴丸弘昭, 吉田将, 「日英機械翻訳のための述部処理システム」電子通信学会論文誌 '77/10 Vol. J60-D №10
- ③ 首藤公昭, 檜原斗志子, 吉田将, 「日本語の機械処理のための文節構造モデル」電子通信学会論文誌 '79/12 Vol. 62-D №12
- ④ 首藤公昭, 吉村賢治, 津田健蔵, 「日本語技術文における並列構造と文の概形」情報処理学会自然言語処理技術シンポジウム 昭和59年11月
- ⑤ 吉村賢治「専門用語の自動抽出」福岡大学工学集報第35号 昭和60年9月

4.5.3 国内調査結果の関連図表

ここでは、国内調査の一貫として実施した「自然言語解析用辞書アンケート調査」への回答内容について、図表としてまとめ、以下の回答内容の分類により一括して掲載した。

- 表Ⅰ プロジェクト、および、システムの概要
- 表Ⅱ 機械可読用辞書について
- 表Ⅲ 辞書の作成、および、メンテナンスについて
- 表Ⅳ 辞書システムの運用
- 表Ⅴ 共用使用について

なお、各表中に「Fig. n-m参照」とあるのは、アンケート回答に記載されていた図表につけた参照番号であり、138 ～ 178 ページに掲載してある。

I、プロジェクト、および、システムの概要(I-1)

調査対象機関	(1)電総研 新明解	(3)京都大学 (長尾研) ニュー コンサイス	(4)京都大学 (長尾研) LONGMAN	(5)インタープレ ス (用語ファイル)	(8)東洋情報 電字林
プロジェクト名	自然言語理解シ ステムの研究 (特別研究)	一般辞書に付き これを直接使用 しているプロ ジェクトはなし	一般辞書に付き これを直接使用 しているプロ ジェクトはなし		電字林(ツール)
システム名	特になし				電字林
目的	実験用自然言語 処理・データベ ース検索				商用電子辞書検 索
使用計算機	FACOM380, DEC2060	FACOM M382	FACOM M382		NEC PC9800シ リーズ
OS	MSP-F4, TOPS20	OS IV/F4 MSP	OS IV/F4 MSP		MS-DOS
使用言語	FORTRAN, Standard LISP	PL/I	PL/I, AIM/RDB		C,アセンブラ
システムの 規模(データ量)	20MB				25MB
システムの規 模(プログラム)	1MB				100KB
機能図	Fig.1-1参照				Fig.1-2参照

I、プロジェクト、および、システムの概要(I-2)

調査対象機関	(8)東洋情報 電字林	(8)東洋情報 電字林	(9)Mu 科学技術 庁辞訳 プロジェクト	(10)ICOT 電子辞書 システム	(11)IPA 日本語基本動詞 辞書
プロジェクト名	(マスターの開発)	(マスターの開発)	日英科学技術文 献の速報システ ムに関する研究	自然言語処理用 電子化辞書	計算機理解のた めの日本語辞書 調査
システム名			日英科学技術文 献の速報システ ム	特になし	特になし
目的			自動翻訳		(注)辞書のみを つくるプロジェ クトでシステム はつくらない。
使用計算機	FACOM M380- R		FACOM M380	PSI (Prolog マシ ン)	
OS	OS IV		FACOM OSIV/F4 MSP	SIMFOS	
使用言語			UTILISP, PL/I	ESF (Prolog + object 指向)	
システムの 規模(データ量)	30-40MB	152MB	文法(3000 ルー ル)、辞書(21万 語)		
システムの規 模(プログラム)			約20,000ステッ プ(8MB)		
機能図			Fig.1-1参照		

I、プロジェクト、および、システムの概要(I-3)

調査対象機関	(14)JAPIO 特許情報機構	(15)日本電気	(16)富士通	(16)富士通	(16)富士通
プロジェクト名		機械翻訳	ATLAS-I	ATLAS-II	ATLAS-I ATLAS-II 科学技術用辞書
システム名	PATOLIS	PIVOT	ATLAS-I	自動翻訳システム ATLASII	ATLAS-I ATLAS-II
目的	商用検索	商用システム	商用自動翻訳	商用自動翻訳	商用自動翻訳
使用計算機	HITAC	ACOS	FACOM Mシリーズ S3000シリーズ	FACOM Mシリーズ	FACOM Mシリーズ S3000シリーズ
OS	VOSIII	ACOS 4, ACOS6	OSN/F4 MSP OSN/X8 FSP OSIS/S		
使用言語	アセンブラ	C言語	アセンブラ、 FORTRAN77	システムプログラム言語	アセンブラ、 FORTRAN77
システムの 規模(データ量)	31GB	4MB		30MB	
システムの規模 (プログラム)	60KB	800KB		100Kstep	
機能図	Fig.1-4参照		Fig.1-5参照	Fig.1-6参照	

I、プロジェクト、および、システムの概要(I-4)

調査対象機関	(17)東芝	(18)ブラビス	(20)日立	(20)日立	(22)日本IBM
プロジェクト名	機械翻訳システム	日英機械翻訳システム 開発プロジェクト	機械翻訳システムの研究	機械翻訳システムの研究	英日機械翻訳
システム名	特になし	ブラビス・バック	ATHENE/E	ATHENE/N	SHALT
目的	商用自動翻訳	商用自動翻訳	翻訳支援	翻訳支援	計算機マニュアルの英日機械翻訳実験システム
使用計算機	VX-700	VAX11, PDP11, PC98シリーズ IBM5550/5560	HITAC Mシリーズ	HITAC Mシリーズ	IBM3081
OS	UNZX	VMS, RSX11-Mplus, MS-DOS	VOS3	VOS3	VM
使用言語	C	Fortran, C, アセンブラ	PL/I	文法記述用言語, PL/I	LISP/VM
システムの規模(データ量)		約15MB			
システムの規模(プログラム)		約1MB			
機能図	Fig.1-7参照	Fig.1-8参照	Fig.1-9参照		Fig.1-10参照

I、プロジェクト、および、システムの概要(I-5)

調査対象機関	(23)シストラン	(24)KDD	(25)NTT	(26)沖電気
プロジェクト名	シストラン自動 翻訳 システム開発プ ロジェクト	機械翻訳	LUTE	機械翻訳システ ム
システム名	シストラン自動 翻訳システム	VALANTINE, KATE	LUTE	未定
目的	商用自動翻訳シ ステム	実験機械翻訳シ ステム	実験機械翻訳シ ステム	商用機械翻訳シ ステム
使用計算機	IBM360/370/ 3033/3084/4341 及びIBM互換機	Symbolics 3600	Symbolics Lisp Machine	if1000
OS	VS1,VS2, MVS	Symbolics リリース60		UNIX
使用言語	シストランマク ロ言語	Lisp		C言語
システムの 規模(データ量)	英日 4MB 日英 4MB	1.5M Byte		
システムの規 模(プログラム)	システム制御プ ログラム2.3MB	7000		
機能図	Fig.1-11参照			

II、機械可読用辞書について(II-1-1)

調査対象機関	(1)電総研 新明解	(3)京都大学 (長尾研) ニュー コンサイス	(4)京都大学 (長尾研) LONGMAN	(5)インタープレス (用器ファイル)
辞書名	新明解国語辞典	ニューコンサイス英和辞典	Longman Dictionary of Contemporary English	IP電子辞典シリーズ(現在5点作成済み)
作成日	SISコード版:昭和54年 JIS版:昭和57年 第3版:昭和60年			昭和58年4月より各種IP電子辞典作成
種類	既存の冊子体辞書の機械可読化	既存の冊子体辞書の機械可読化	既存の冊子体辞書の機械可読化	既存の冊子体辞書の機械可読化 翻訳システムの解析用辞書
種別	日本語	和英	英和	英和
対象分野 一般分野	国語辞典 (約58,000)	一般	一般	
専門分野				科学技術用語データベース 収録単語数 250,000 ビジネス用語データベース 収録単語数 100,000 科学技術活用データベース 文例 1,400 科学技術熟語表現データベース 文例 3,000 科学技術略語データベース 収録単語数 40,000
今後の予定	今後元の辞書の改定があればそれに沿った変更を行う予定であるが今のところ特に予定はない。	京都大学における追加予定なし (辞書の性格上辞書内容の変更は冊子辞書の著者及び三省堂のみが可能である。)	京都大学における追加予定なし	科学技術関係分野の用語を中心に50万語を収録予定で作業を継続中 科学技術用語データベース 予定件数 500,000 ビジネス用語データベース 予定件数 200,000 科学技術略語データベース 予定件数 60,000

II、機械可読用辞書について(II-1-2)

調査対象機関	(8)東洋情報 電字林	(8)東洋情報 電字林	(8)東洋情報 電字林	(9)Mu. 科学技術 庁翻訳 プロジェクト (英語)	(9)Mu. 科学技術 庁翻訳 プロジェクト (英和)
辞書名	電字林対訳辞書	電字林基本語辞 書マスター	電字林基本語辞 書マスター	英語辞書MTF	英日変換辞書 MTF
作成日	84年8月～ 86年3月	84年10月～ 85年11月	84年9月～ 86年4月	1986年3月	1986年3月
種類	対訳電子辞書	新規編集 (既存辞書用語辞 典等を参考)	既存、機械可読 化辞書のフォー マット変更辞書	機械翻訳システ ムの解析生成用 辞書	英日変換用辞書
種別	和英および 英和	和英および 英和	かな、日本語、 英語の対訳辞書	英語	英和
対象分野 一般	ビジネス用途 和英10万語 英和10万語	ビジネス用途 和英10万語 英和10万語			
専門	Fig. II-1-1		Fig. II-1-1の内 情報処理 電気・電子 ビジネス プラント 略 語 土木・建築 軍 事	電気(20,000語) その他 (50,000語) 学術用語	電気
今後の予定	医学分野10万語 日中英対訳辞書 を86年度に予定	未定	日中英辞書	プロジェクト完 了時点の1986年 3月には専門用語 10万語予定して いる。	プロジェクト完 了時点の1986年 3月には2万6千 語を作成する予 定である。

II、機械可読用辞書について(II-1-3)

調査対象機関	(9)Mu 科学技術 庁翻訳 プロジェクト (日本語)	(9)Mu 科学技術 庁翻訳 プロジェクト (和英)	(10)ICOT 電子辞書 システム	(11)IPA 日本語基本動詞 辞書	(14)JAPIC 特許情報機構
辞書名	日本語辞書MTF	日英変換辞書 MTF	基本語電子化意 味辞書	計算機用 日本語基本動詞 辞書	PATOLIS フリーキーワー ド一覧
作成日	1986年3月	1986年3月	1984年6月 ～1986年(3ヶ年)	昭和59年4月～ 昭和61年3月	第2版: 昭和58年10月
種類	機械翻訳シス テムの解析・生成用 辞書	日英変換用辞書		日本語処理用の 汎用辞書	検索用語集およ び蓄積用
種別	日本語	和英	日本語、和英 英和、英語	日本語の基本動 詞	日本語
対象分野 一般			国語 6,000 英語 2～3,000 和英 2～3,000 英和 2～3,000	特に分野は限ら ない。(約870)	
専門		電気(20,000語) その他 (50,000語) 学術用語			全産部門 (約330,000語)
今後の予定	プロジェクト完 了時点の1986年 3月には専門用語 10万語予定して いる。	プロジェクト完 了時点の1986年 3月には専門用語 10万語予定して いる。	国語辞書につい ては、6万語ま で拡張予定	収録済みのもの に次いで重要な 和語動詞および 漢語や変動詞を 今後収録してい く予定である。	用語整理を行っ ている。

II、機械可読用辞書について(II-1-4)

調査対象機関	(15)日本電気	(15)日本電気	(15)日本電気	(15)日本電気	(15)日本電気
辞書名	LEXICON	マスター辞書	英和辞書	AI用語主集	NEC専門用語集
作成日				60年	
種類	機械翻訳	人手によってあつめた。	冊子辞書の機械化物	解析、生成用	機械翻訳
種別	日本語/英語	日本語	日本語、英和	日英	日英
対象分野 一般分野	収録単語数 20,000	収録単語数 89,000	収録単語数 約13万		
専門分野	電子機器 収録単語数 15,000 その他 収録単語数 10,000			AI 収録単語数 6000	情報処理 収録単語数 25,000
今後の予定	一般分野予定件数 40,000 電子機器予定件数 20,000 その他 予定件数 20,000				情報処理 予定件数 40,000

II、機械可読用辞書について(II-1-5)

調査対象機関	(16)富士通	(16)富士通	(16)富士通	(17)東芝	(18)ブラビス
辞書名	ATLAS-I基本辞書	日本語辞書(英語辞書)	ATLAS-I ATLAS-II 科学技術用語辞書	特になし	日英機械翻訳辞書
作成日		現在も更新中		1978～	昭和60年8月 (約2ヶ年)
種類	機械翻訳システムの解析 生成用辞書	機械翻訳システムの生成 解析用辞書	専門用語辞書	機械翻訳用 構文・意味解析辞書	機械翻訳システムの解析用辞書
種別	英日	日本語(英語)	ATLAS-I:英日 ATLAS-II:日英	英和、和英	和英
対象分野 一般	(53,000語)	科学技術(5万語)		一般(30,000語)	(約60,000語)
専門			科学技術分野 300分野 (約25万語)	(250,000語)	
今後の予定		現在ある単語の 精度を高めること に力を注ぐ。			

Ⅱ, 機械可読用辞書について(Ⅱ-1-6)

調査対象機関	(20)日立	(20)日立	(22)日本IBM	(23)シストラン
辞書名	ATENE/E辞書	ATENE/J辞書		シストラン辞書データベース
作成日		作成中	1980年より 継続中	
種類		機械翻訳用辞書	機械翻訳システムの英日変換用 名詞辞書	シストランシステムの 解析及び生成辞書
種別		日英	英和	日英 英日
対象分野 一般		科学技術文献		
専門	経済		コンピューターマニュアル に現れる名詞及び複合名詞、 収録単語数は 11424	
今後の予定			最終的に必要になると思われる 語数について現在検討中である。	基本辞書—需要の高い分野を優先的に広範な分野をカバーするような網羅的な辞書をめざす。 ルール付き表現辞書基本辞書で与えられる意味以外の訳しわけの対応が目的であるがキメ細かい翻訳を得るには、拡充が必須であり重要課題の重大保大のひとつとして予定が組まれている。

II、機械可読用辞書について(II-1-7)

調査対象機関	(24)KDD	(25)NTT	(26)沖電気
辞書名	なし		なし
作成日	1.5年 (作成期間)		約6ヶ月
種類	機械翻訳システム の解析		機械翻訳システム の解析
種別	英和	日本語、和英 英和、英語	和英 英和、英語
対象分野 一般	ニュース	各数万以上	主に科学技術分 野(約50,000)
専門			
今後の予定			

II、機械可読辞書について(II-2-1)

調査対象機関	(1)電総研 新明解	(3)京都大学 (長尾研) ニュー コンサイス	(4)京都大学 (長尾研) LONGMAN	(5)インタープレ ス (用語ファイル)	(8)東洋情報 電字林
データ項目	データは1レ コード64文字 (128バイト)で各 見出し語ごとに 改行する他は、 ほぼ元の国語辞 典通りのベタ書 きになっている。 各レコード 内でのデータの 区別は下のとう りである。 Fig.II-2-1参照	資料 Fig.II-2-3参照	資料 Fig.II-2-5参照	資料 Fig.II-2-18参照	[日英] (専門用語) かな...検索のためのキー 日本語...かなに対応した日 本語見出し 英語...日本語の対訳 (基本語) かな...検索のためのキー 日本語...かなにグループ化 された日本語見出し 英語...日本語の対訳 [日英] (専門用語) 英語...検索のためのキー 日本語...英語の対訳 (基本語) 英語...検索のためのキー 日本語...英語の対訳
データ項目の 内容構成	資料 Fig.II-2-2参照	資料 Fig.II-2-4参照	意味マーカの階 層構造 資料 Fig.II-2-6参照 Fig.II-2-7参照		データ項目に記述済み

II、機械可読辞書について(II-2-2)

調査対象機関	(8)東洋情報 電字林	(8)東洋情報 電字林	(9)Mu 科学技術 庁翻訳 プロジェクト	(10)ICOT 電子辞書 システム
データ項目	[和英] 語番号 かな ...日本語の 読み 日本語 ...必要に 応じて分野およ び意味を付加 英語 ...日本語に 対する対訳 [英和] 語番号 英語 品詞 ...必要最小 限 日本語 ...英語の 対訳	(1)かな (2)日本語 (3)英語 全て1対1対応 かなは日本語の 読み、ただし日 本語には一部注 釈がある。日本 語の中の数字及 び英大文字はか なとして採用し ている。	資料 日本語品詞一覧 Fig.II-2-8 英語品詞一覧 Fig.II-2-9 日本語格ラベル 一覧 Fig.II-2-10 英語格ラベル一 覧 Fig.II-2-11 意味マーカ体系 表 Fig.II-2-12	日本語辞書 (1)見出し語・ID番号 ・かな表現 (2)品詞情報 ・見出し語 (原型、変化型) ・発音 ・文法情報 ・関連語のリスト (タイプ、ID番号、関連語) ・補足説明 (分野、用例記号、その他) ・意味情報 (意味、用例、文法情報、文 型、イディオム情報) ・特別な意味に対する関連 (語のリスト、タイプ、 ID番号、関連語) ・特別な意味に対する補足説明 (分野、用法、その他)
データ項目の 内容構成	対訳関係		上記資料参照	

II)、機械可読辞書について(II-2-3)

調査対象機関	(11)IPA 日本語基本動詞 辞書	(14)JAPIO 特許情報機構	(15)日本電気	(15)日本電気	(15)日本電気
データ項目	資料 Fig. II-2-13		語基区切り 形態素接続 品詞 シソーラスコード 格パターン 通語パターン など約60種類 いつもこれらが 全てのエントリーについてい るわけでない。	表記、 品詞、 意味コード Articulation point、 読み、 頻度	冊子辞書内容を 形式化したもの、 例文つき
データ項目の 内容構成	前記のデータ項目のうち、本辞書IPAの特長ある項である。 (1)述語集 (2)ヴォイス (3)テンス・アスペクト (4)ムード についてその概要をFig. II-2-13'に述べる。				

II、機械可読辞書について(II-2-4)

調査対象機関	(15)日本電気	(15)日本電気	(16)富士通	(16)富士通
データ項目	体言のみ、 意味コード、 対訳、 品詞細分類、 読み、 見出し、 対訳	接続、 読み、 頻度、 格パターン、 表層格パターン、 シソーラスコード、 品詞コード、 など20種類	見出し語…英語の綴りを示す。この辞書をひくときのキーになる。 品詞…品詞名を示す記号 主カテゴリ…各品詞の变化形を示す記号 副カテゴリ…各語の構文特徴、意味分類などを示す記号 標準訳語…標準的に使われる訳語 副訳語…副カテゴリに対応する訳語 Fig.II-2-14参照	Fig.II-2-15参照 Fig.II-2-16参照
データ項目の内容構成				

II、機械可読辞書について(II-2-5)

調査対象機関	(16)富士通	(17)東芝	(18)ブラヒス	(20)日立
データ項目	ATLAS-I, ATLAS-IIの 各基本辞書に準 ずる。	見出し ...英語(英和)、漢字か な混じり(和英) 読み(和英のみ) ...かな 品詞 ...構文分析に用いる統語 カテゴリ 数 ...single,plural,don'tcare 性 ...male,female,don'tcare 人称 ...1,2,3 不規則変化標識 ...不規則変化 する語につく 慣用句標識 ...慣用句につく 意味標識 ...意味分析の一部に 用いる意味カテゴリ 意味分析規則 ...語い遷移文法 (東芝の開発した統語・意味分 析文法)の語い文法部) 格フレーム ...意味分析の一部 に用いる 訳語 ...訳語	ソース語 ...仮名漢字混じ りで表記された日本語 JIS漢字コード ホモグラフ ...ソース語 が1つに対して2つ以上 の品詞の訳語を持つ場 合、そのID 訳語 ...ソース語に対す る訳語(ホモグラフの場 合、それぞれの品詞に 対応する訳語) 属性 ...ソース語、訳語そ れぞれに対する属性情報 活用 ...訳語の活用ID クロス・リファレン ス ...活用(名詞の単複、 動詞の時制)処理のため のソース語と訳語との 関係付け情報	
データ項目の 内容構成	ATLAS-I, ATLAS-IIの 各基本辞書に準 ずる。		公開せず	

II、機械可読辞書について(II-2-6)

調査対象機関	(22)日本IBM	(23)シストラン	(24)KDD	(25)NTT	(26)沖電気
データ項目	1、見出し語:英語の名詞、ないしは、複合名詞 2、意味マーカ:訳し分けに使われる、名詞の意味上の分類を示す。 3、訳語:見出し語に対応する日本語。意味マーカごとに集めてあり、同じ意味マーカに対応するものは、頻度順に並んでいる。	[基本辞書] (1)ソース言語 ソース言語用語情報 形態素解析情報 構文解析情報 意味解析情報 (2)ターゲット言語 対応ターゲット言語 用語情報 形態素生成情報 構文生成情報 意味生成情報 その他 [ルール付き表現辞書] ルール付表現記述 *必要に応じて基本辞書の(1)(2)と同様の情報を付加することができる。			
データ項目の内容構成	意味マーカ一覧 Fig.II-2-17参照				

II、機械可読辞書について(II-3-1)

調査対象機関	(1)電経研 新明解	(3)京都大学 (長尾研) ニュー コンサイス	(4)京都大学 (長尾研) LONGMAN	(5)インタープレ ス (用語ファイル)	(8)東洋情報 電字林
データ項目 のデータフォー マット	Fig.II-3-1参照	Fig.II-3-2参照		Fig.II-2-18参照	現在は公開不可
辞書のファイル フォーマット	磁気テープ上の フォーマットは Fig.II-3-1参照 FACOM M-380 上では 12800バイト固 定長のランダム ファイル(100レ コード分収容)	Fig.II-3-3参照	Fig.II-3-4参照	Fig.II-2-18参照	現在のところ公 開不可 ただし論理構造 はデータ項目に 記述済み

II、機械可読辞書について(II-3-2)

調査対象機関	(8)東洋情報 電字林	(8)東洋情報 電字林	(9)Mu 科学技術 片翻訳 プロジェクト	(10)ICOT 電子辞書 システム
データ項目 のデータフォー マット	各データは全てJEFコード <和英> (1)語番号部:10バイト(5文字) (2)かな部:24バイト(12文字) (2.5)日本語コーク部:2バイト (1文字) (3)日本語部:34バイト(17文字) (3.5)英語コーク部:2バイト (1文字) (4)英語部:78バイト(39文字) <英和> (1)語番号部:10バイト(5文字) (2)英語部:40バイト(20文字) (3)品詞部:4バイト(2文字) (3.5)日本語コーク部:2バイト (1文字) (4)日本語部:74バイト(37文字)	各データは全て JEFコード かな部:100バイト (50文字) 日本語部:80バイト (40文字) 英語部:200バイト (100文字) (原則として空項目 はなし)	Fig.II-3-5参照	
辞書のファイル フォーマット	日本語と英語は対訳の関係 和英 RECFM(FB) LRECL(152) 150+2(シフトコード) 英和 RECFM(FB) LRECL(132) 130+2(シフトコード) (いづれも修正作業用マス ター)	RECFM(F・B) LRECL(380) BLKSIZE(3800) DSORG(PS)	Fig.II-3-6参照	

II)、機械可読辞書について(II-3-3)

調査対象機関	(11)IPA 日本語基本動詞辞書	(14)JAPIO 特許情報機構	(15)日本電気	(16)富士通
データ項目 のデータフォー マット				
辞書のファイル フォーマット	本辞書はユーザには、順編成ファイルの形式で提供する。一つの見出しに対して前述の記述項目が1レコードにおさめられている。 レコード形式:VB(可変長ブロック化レコード) レコード長:8000Byte(最大長) ブロック長:16000Byte(最大長) コード系:JIS第1水準、第2水準漢字コード レコード数:約3200			データ項目間の論理構造についてはFig.II-2-15を参照

II、機械可読辞書について(II-3-4)

調査対象機関	(16)富士通	(16)富士通	(17)東芝	(18)ブラビス	(20)日立
データ項目 のデータフォー マット		ATLAS-I, ATLAS-IIの 各基本辞書に準 ずる。		Fig.II-3-7	
辞書のファイル フォーマット		ATLAS-I, ATLAS-IIの各基 本辞書に準ず る。	各種フォーマッ トが可能であ る。最も単純な フォーマットと して可変長文字 列データとして 構成されてい る。 必要に応じて変 換可能	Fig.II-3-8	

II)、機械可読辞書について(II-3-5)

調査対象機関	(22)日本IBM	(23)シストラン	(24)KDD	(25)NTT	(26)沖電気
データ項目 のデータフォー マット	1.見出し語 ...EBCDIC コー ド可変長 2.意味マーカー ...EBCDIC コー ド 2 バイト (2 文 字) 3.訳語 ...IBM 漢字 コー ド可変長				
辞書のファイル フォーマット	可変長のCMS ファイルであ る。1 行に辞書 レコードが一つ 入っている。各 辞書レコード は、LISP形式の リストである。 (見出し語・訳語情 報リスト) 訳語情報リスト: (訳語情報...訳語 情報) 訳語情報: (意味マーカー... 訳語...訳語)				

III、辞書の作成、および、メンテナンスについて(III-1-1)

調査対象機関	(1)電総研 新明解	(3)京都大学 (長尾研) ニュー コンサイス	(4)京都大学 (長尾研) LONGMAN	(5)インタープレス (用語ファイル)
作成、メンテナ ンス体制	出版社と共同	三省堂がメンテナ ンスを行って いる。	英国Longman社 が入力	自社においてすべて行う。
作成体制				
人員構成	情報処理技術者 1名 外注 1名 入力 外注			
工数	24人日			
経費	インプットから含め ると数千円			
メンテナンス 体制				
メンテナンス 内容	メンテナンスは、主 として元の辞書(本の 形)と比較しながら、 インプット時の誤り をチェックする。つ まり、一般の校正と 同じ仕事を行う。		(京大におけるメ ンテナンスはな し)	誤植等入力ミスのチェック・ 修正新規用語の追加
頻度	版改訂のたびごと			常時
手順	メンテナンス用に開 発した専用のエディ タを用いて校正作業 を行う。最近では日本 語端末上でFACOMの 日本語エディタを用 いて校正も行ってい る。			[誤植等の見直し] (1)全件リストによるデータの校正 (2)校正済みリストのデータパンチ (3)再校正 (4)媒体変換(FD→MT) [新規用語の追加] (1)用語データの作成 (2)データパンチ (3)校正 (4)再データパンチ (5)ソースデータベースとのマッ チング、ソート (6)新データベース作成完了
人員構成	外部からの専任 オペ レータ 1名			翻訳者 4名 オペレータ 7名 編集者 3名
経費	年100~200万程度			

III、辞書の作成、および、メンテナンスについて(III-1-2)

調査対象機関	(8)東洋情報 電字林	(8)東洋情報 電字林	(8)東洋情報 電字林
作成、メンテ ナンス体制	社内にてパソコン用DB化	入力編集はアイディ社と共同	入力、修正(フォーマット変換を含む)はインタープレス社と共同
作成体制			
人員構成	オペレータ1名	データ作成・編集(辞書編集経験者・米英在住経験者)10名 インプットオペレータ(追加・編集のみ)1名 データのMT化 外注	＜修正とフォーマット変換のみ＞ インプットオペレータ 1名
工数	一人月(100万語)	130人月	10人月
経費		4-5000万円(現時点で、CPU費込み)	300-500万円 (計算機費込み)
メンテナンス体制			
メンテナンス 内容		各種エラーの修正・追加	入力データはJISであるためJISかコードらJEFのコード変更を行う。 データ容量圧縮のため、一部データ変更をおこなう。 作業上発生したエラーの修正
頻度		随時	現在は随時作成段階 将来は半年に一度を予定
手順		メンテナンス必要時に、MTからloading 保管時にはMTにunloading	(1)インタープレスより入手のオリジナルMTをM380Rにローディング (2)コード変換 (3)エディット用フォーマット変換 (4)PFDによる修正 (5)マスターへのreformat (6)ソート処理
人員構成		ブルーフリーダー・インプットオペレーター 2名	インプットオペレータ 1名
経費		数十万円/月	

III、辞書の作成、および、メンテナンスについて(III-1-3)

調査対象機関	(9)Mu. 科学技術庁翻訳プロジェクト	(10)ICOT 電子辞書システム	(11)IPA 日本語基本動詞辞書
作成、メンテナンス体制	電総研、JICST、京大が協力して作成。データは外注	辞書出版社と共同 英語(富士通、旺文社)、英日(東芝、学研)、日本語(日立、未定)、日英(富士通、旺文社)	大学・研究所の国語学・言語学の研究者によるワーキンググループを組織
作成体制			
人員構成	辞書作業分科会17名 辞書WG 25名 機関 10名 外注 50名以上		国語学・言語学専門家 17名 情報処理技術者 3名 プログラマ 2名 アルバイト 3名
工数			約8人月
経費			
メンテナンス体制			
メンテナンス内容	(1)一元管理 辞書ファイルを汎用のデータベースシステム上で構築し、定期的に更新、追加を行う。 (2)機械翻訳総合システムにおいて、ユーザが単語単位で更新、追加を行う。	実験用で、しかも規模が小さいのでユーザにまかせている。	(1)各記述項目の記入ミス (2)言語学・国語学の研究者による辞書内容の校閲 (3)新しい単語の登録
頻度	(1)数ヶ月に一回(大量) (2)常時(私用辞書)		
手順	(1)大量のデータを外注で記入フォーマットにより作成、一元管理システムによりDB管理が行われる。外部へはLISPS式でMTFにより提供される。 (2)総合システムの辞書エディタサブシステムを用いて、機械翻訳用の辞書ファイル(MTF,VSAM)及び内部辞書DBの編集が行われる。ユーザは私用辞書DBの編集がおこなわれる。ユーザは私用辞書に対してのみ編集が可能であろう。		
人員構成	(1)未定 (2)システムは自動化されている。ユーザの作業		作成体制と同じ
経費	(1)未定 (2)計算機使用料VSAM使用料		

III、辞書の作成、および、メンテナンスについて(III-1-4)

調査対象機関	(14)JAPFO 特許情報機構	(15)日本電気	(15)日本電気	(15)日本電気	(16)日本電気
作成、メンテナ ンス体制	用語の整理は ユーザの協力を 得ている	社内及び関係会 社	社内及び関係会 社	社内及び関係会 社	
作成体制					
人員構成					
工数					
経費					
メンテナ ンス体制					
メンテナンス 内容	キーワードの表 記統一 同義語の整理	追加、更新	現在なし	エラー発見時の 訂正	
頻度		必要発生時			適宜追加
手順		入力および更新 項目、更新関連 候補の抽出をお こなう。			
人員構成	外部協力者を含 め約50名				
経費					

III、辞書の作成、および、メンテナンスについて(III-1-5)

調査対象機関	(15)日本電気	(16)富士通	(16)富士通	(16)富士通	(17)東芝
作成、メンテナンス体制	社内及び関係会社	独自開発	入力 は FIE と共同	インクプレス社"IP電子辞書"のデータを元にして属性追加等の加工を加えた。	
作成体制					
人員構成					
工数					
経費					
メンテナンス体制					
メンテナンス内容	追加・誤り訂正	基本辞書としての枠内での見出し語の補充		誤りの訂正	
頻度	現在まで3回	3ヶ月に一度		随時	
手順	新単語登録要求抽出	通常の製品登録、出荷手順による。	Fig. III-1-1参照	通常の製品登録、出荷手順による。	
人員構成					
経費					

III、辞書の作成、および、メンテナンスについて(III-1-6)

調査対象機関	(18)ブラビス	(20)日立	(22)日本IBM	(23)シストラン
作成、メンテナンス体制	当社要員が専任	社内開発	単独	
作成体制				
人員構成	公開せず		情報処理技術者1名 データの収集・分類補助員 2名 インプット・オペレータ 外注	
工数				
経費				
メンテナンス体制				
メンテナンス内容	登録語の追加 既登録語の属性情報の誤り修正 訳語の適切チェック		新規レコードの追加 既存レコードの修正 意味マークの訂正	
頻度	6ヶ月-1ヶ年		約10レコード/週	
手順	1、追加登録すべき語の洗い出し 2、属性情報の誤り箇所洗い出し 3、入力 4、チェック		1、翻訳結果のチェックにより未登録語及び不適切な訳語や意味マークを発見する。 2、その語が、含まれる例文を例文ファイルから検索する。その結果と1の原文を考慮して訳語や意味マークを追加、ないしは修正する。	
人員構成	公開せず		情報処理技術者1名	
経費	公開せず			

III、辞書の作成、および、メンテナンスについて(III-1-7)

調査対象機関	(24)KDD	(25)NTT	(26)沖電気
作成、メンテナンス体制	辞書作成会社と共同	なし	
作成体制			
人員構成	言語学専門家 2名 情報処理技術者 1名		
工数			
経費		0	
メンテナンス体制			
メンテナンス内容		なし	
頻度			
手順			
人員構成			
経費			

III、辞書の作成について(III-2-1)

調査対象機関	(1)電総研 新明解	(3)京都大学 (長尾研) ニュー コンサイス	(4)京都大学 (長尾研) LONGMAN	(5)インタープレス (用語ファイル)
辞書の作成手順	1.元の辞書を渡してインプットを外注 2.磁気テープで納入 3.ディスク上で修正		データの作成、入力 は英国Longman社が行ったものであり京都大学における作成作業はなし。 京都大学ではそのデータを使用目的(たとえば英文のパーサ言語資料の作成に適した形式にプログラム(主としてPL/I)により変換して使用している。	(翻訳等原データ作成) (1)プログラム作成 (2)入力画面作成 (3)入力(フロッピーディスク) (4)出力 (5)校正 (6)媒体変換(フロッピーディスクから磁気テープへ)
単語の収集方法	既製の辞書を用いているので特に収集作業は行っていない	既存の冊子体辞書の機械可読化であり三省堂ニューコンサイス英和辞書の全単語を入力。	既存冊子体辞書を機械可読化したものでありLongman Dictionary of Contemporary Englishの全単語が入力されている。	(1)従来の当社で発行した月刊誌または単行本からの抽出 (2)英文は米国軍用規格、民間企画のマニュアル等から抽出し翻訳 (3)公的機関から発行された用語等 (4)市販各種辞典からの選択・抽出
収録単語の分類 作業の手順	特に分類は行っていない			(1)用語収集(用語検討、用語統一、訳語作成・選定) (2)用語データバンチ(日本語—日本語の読み—英語)
収録単語の選定 基準				

III、辞書の作成について(III-2-2)

調査対象機関	(8)東洋情報 電字林	(8)東洋情報 電字林	(8)東洋情報 電字林	(9)Mu 科学技術庁翻訳 プロジェクト
辞書の作成手順	汎用計算機上の 辞書マスターを 汎用計算機上で パソコン用DB化 (圧縮用の特殊 コード使用) その後転送、パ ソコン上にてイ ンデックス等作 成	(1)データ用紙記入 (2)MT成作 (外注) (3)チェックプログラム及 び目によるエラー摘出 (4)PFDによる修正 (5)ソート処置		Fig.III-2-1参照
単語の収集方法	別回答参照	(1)ビジネスユース上基本 語(見出し語)を既存の辞書 から選出 (2)(株)アイデンの蓄積デー タを中心として翻訳を当て はめる。 (3)既存の辞書、用語辞典を 参考にして不備な点を補正		(1)INSPRC抄録3000文に出現した英単 語 (2)日英変換辞書の対訳語として登録さ れた英単語 (3)名詞・動詞以外の基本単語
収録単語の分類 作業の手順		単語の収集については単語 の収集方法を参照		(1)単語の収集及び各種マーカ付与に KWICが用いられている。 (2)体言と用言に分かれていて作成作業 が行われている。 (3)名詞に対する各種マーカは辞書や用 語集を用いて行っている。 (4)用語に対する各辞書項目の付与は KWIC等出現した文を中心に分析を行い 付与作業を行う。 (5)いずれの場合も詳細な作業マニユ アルが作成されている。 (6)付与作業は外注で行われているが言 語的知識を持つ人間が担当している。
収録単語の選定 基準		ビジネスユースを主体に 選定 専門用語を連接するのに役 立つ語の選定		

III)、辞書の作成について(III-2-3)

調査対象機関	(10)ICOT 電子辞書 システム	(11)IPA 日本語基本動詞辞書	(14)JAPIO 特許情報機構	(15)日本電気
辞書の作成手順	1 リサーチグループで基本仕様書(Logical format)を作成 2 その仕様に基づいて、言語データを収集・分析 3 入力 4 修正	1)辞書の記述の枠組み及び収録語の決定 2)記述マニュアルの作成 3)マニュアルに従って各単語に対し、記述シートに各項目内容を記述 4)記述内容の校閲 5)記述シートをVSAM形式の辞書ファイルにツールを用いて増末を介して入力	Fig.III-2-2	1. 単語抽出をおこなう。 2. データフォームに記入し、一括登録をおこなうか、或いはエディタ登録更新をおこなう。
単語の収集方法	一般の辞書を参考に、最重要単語を収集する。	国立国語研究所の種々の調査結果及び種々の市販辞書等から収集		市販辞書、テキストの中から単語を抽出し、形態解析による未定義語推定をおこなう。 辞書コンサルト
収録単語の分類作業の手順				1. 分類表に基づいて品詞分類をおこなう。 2. 分類表に基づいて意味分類をおこなう。 3. 辞書データ手順書に基づいて辞書コーディングをおこなう。
収録単語の選定基準	いわゆる最重要語	意味的にも使用頻度の上からも語彙体系上最も重要と考えられる動詞を、国立国語研究所の種々の調査及び市販辞書の重要語の選択を参考にして選んだ。	一件の公開特許抄録文中(約300文字)の技術用語のうち重要な語を約15ほど選定している。	

III、辞書の作成について(III-2-4)

調査対象機関	(15)日本電気	(15)日本電気	(15)日本電気	(15)日本電気	(16)富士通
辞書の作成手順	1、一般辞書の見出しを入力する。 2、大量テキストを入力し、単語を抽出し、頻度付けをおこなう。 3、1、2、の結果をフォーマットに記述する。 4、登録する。 5、verifyする。		市販専門用語学、テキストから抽出し、データフォームに記入する。	1、データフォーマット作成 2、ワープロ入力 3、ホスト辞書登録	
単語の収集方法		市販辞書		社内技術文書から抽出 市販専門用語集から抽出	市販の小型英和辞典に収録されている見出し語をベースとし、科学技術用として若干の調整を加えた。
収録単語の分類作業の手順			1、意味分類付与 2、見出し語の品詞分類付与および対訳語の品詞分類付与 3、1、2、の結果をデータフォーム記入 4、登録	1、体言、用言の分類 2、体言なら、意味コード付与し、他品詞細分付与、フォーマット記述 3、用言なら、表層格マーカ抽出、深層格マーカ抽出、意味コード付与、フォーマット記述	
収録単語の選定基準	テキスト選定				市販の小型英和辞典に収録されている見出し語をベースとし、科学技術用として若干の調整を加えた。

III、辞書の作成について(III-2-5)

調査対象機関	(16)富士通	(16)富士通	(17)東芝	(18)ブラビス	(20)日立
辞書の作成手順		インタプレス社「IP電子辞典」のデータを元にして属性追加等の加工を加えた。	(1)見出し決定 (2)品詞以下各統語情報の付加 (3)意味分析用各情報の付加 (1)(2)は比較的閉じた情報であるが(3)は開いた情報であり手間がかかる。	Fig.III-2-5参照	
単語の収集方法		インタプレス社「IP電子辞典」に準ずる。	(1)頻度の高い基本語の収集 (2)実データ(文章)から補填	公開せず	
収録単語の分類作業の手順		インタプレス社「IP電子辞典」のデータを元にして属性追加等の加工を加えた。	Fig.III-2-3参照	辞書登録プログラムを作成し、入力の際インプット・オペレータと画面上で会話することにより、各種マークは自動的に付与されるようになっている。	
収録単語の選定基準		インタプレス社「IP電子辞典」に準ずる。	常識的に頻度の高いと思われるもの 本システムは辞書を3分類しておりユーザに特有な用法の語はユーザ辞書に登録する。 Fig.III-2-4参照	一般用語として煩雑に使用されると思われる単語類	

III、辞書の作成について(III-2-6)

調査対象機関	(22)日本IBM	(23)シスト ラン	(24)KDD	(25)NTT	(26)沖電気
辞書の作成手順	1. "IBM情報処理用語英和対訳集"(機械可読)を再編成し、意味マーカを付加する。 2. IBMソフトウェア・マニュアル(英文)に使われている語を調査し、意味マーカ、訳語を付けて1. で作ったデータに加える。				
単語の収集方法	1. "IBM情報処理用語英和対訳集" 2. IBMソフトウェア・マニュアル(英文)から名詞と複合名詞を抜き出す。		基本的なものは辞書類から、その他は雑誌・新聞から。		
収録単語の分類 作業の手順	1. "IBM情報処理用語英和対訳集"の中の名詞、複合名詞に意味マーカを付ける。 2. 英文マニュアル文を単語に分割し、形態素解析を行い、集計する。 3. その中から名詞を抜き出す。 4. もとの英文を参照しつつ意味マーカを付ける。				
収録単語の選定 基準	1. "IBM情報処理用語英和対訳集"に含まれる名詞及び複合名詞は、すべて登録する。 2. 調査した英文マニュアルに出てくる名詞は、全て登録する。但し、大文字のアルファベット数文字からなるソフトウェアの名称(例:DOS,MVS等)は除く。複合名詞については、個々の単語の訳語から全体の訳語が合成できない場合に登録する。				

Ⅲ、辞書の作成について(Ⅲ-3-1)

調査対象機関	(1)電総研 新明解	(3)京都大学 (長尾研) ニュー コンサイス	(4)京都大学 (長尾研) LONGMAN	(5)インタープレ ス (用語ファイル)
管理システムの 機能概要	少し古いエディタ であるが下記のよ うになっている。 現在のものも大体 同じである。 FigIII-3-1参照	メンテナンス用 システムは現在 運用していない (辞書入力時の校 正用ソフトウェ アは開発し使用 した)。ここに は原データから 検索用データ ベースを作成す るので機能概要 を示す。 FigIII-3-2参照	メンテナンス用のシステム はない。検索等利用のために 京都大学で作成したプログラ ムについて記述する。 FigIII-3-3参照	
入力方法 (データ、 エントリー、 ルール)	特になし		不明	
問題点、意見	既製の日本語エ ディタには、作成 当時良いものがな く結局自作する事 になった。しかも データが途中で人 手に渡り、汚いま ま拡散したので、 データの管理体制 をきちんとするこ とも大事であるこ とを痛感した。さ らにきちんとした ドキュメントを 作っておく事が自 分にとっても大切 事が分かった。		LDOCEは計算機処理を前提 として作成された部分が多い ので、比較的情報の抽出等に 利用しやすくなっている。 しかし文法コードの記述形式 の一部に不規則なものがある こと、各項目中に情報を分離 しにくい部分があることなど、 計算機から見て必ずしも 完全なもの(処理しやすいも の)ではない。今後冊子体辞書 と計算機用辞書の接点を考え れば、冊子体辞書上には表れ ないが計算機処理用には利用 できる情報(単純なものとし ては項目分離用コードの挿入 から考えられる)を付加する など、作成時の考慮が必要で あろう。	

III)、辞書の作成について(III-3-2)

調査対象機関	(8)東洋情報 電字林	(8)東洋情報 電字林	(8)東洋情報 電字林	(9)Mu 科学技術庁翻訳 プロジェクト
管理システムの 機能概要	Fig.III-3-4参照	メンテナンス手 順参照	メンテナンス手 順参照	Fig.III-3-5参照
入力方法 (データ、 エントリー、 ルール)	別回答参照	非公開	公開不可	Fig.III-3-6参照
問題点、意見	同上	<問題点> (1)データの構成 編集要領を一過 り習得するのに 約1ヵ月を要す る。 (2)データの追加 には多くの人日 を要する。	中国語辞書の作 成に至った場合 の簡体文字の取 り扱い不可の解 決	(1)一元管理される辞書汎用 D.B上に構築することにより管 理の効率的な適用をはかる。 (2)辞書データ項目については専 門家による委員会を設置し承認 されたもののみを登録する。そ の際対訳を考慮して複合語の単 位を決定する。 (3)修正は定期的に一括して行 う。 (4)システム上で用いるときは 私用辞書を利用させることによ り使い勝手をよくする。またこ の結果得られた辞書を(2)に提出 する。 (5)外部提供のフォーマットを標 準化する。その際辞書項目、構 造を柔軟なものとし選択・追加を 自由に行えるようにする。

III)、辞書の作成について(III-3-3)

調査対象機関	(10)ICOT 電子辞書 システム	(11)IPA 日本語基本動詞辞書	(14)JAPIC 特許情報機構
管理システムの 機能概要	1. 基本仕様書(Logical format)に基づいたテキスト形式の辞書テープを作成する。 2. DAM形式のファイルの基本形式に変換する。 3. 追加、修正をエディターでおこなう。	1)手書きの辞書シートをTSS端末上から効率よく磁気ディスク上の辞書ファイルに入力する。 2)辞書内容の検索 3)入力済辞書内容の修正 4)磁気ディスク上の辞書ファイルの磁気テープへのバックアップ 5)辞書内容の印刷	
入力方法 (データ、 エントリー、 ルール)			
問題点、意見	・ ICOTでは、第一段階では、日本語、和英、英語、英和の開発を独立にすすめている(FIG. III-3-7)。これを基にしてそれぞれの辞書をmergeして、汎用マスター辞書の作成を試みる。 ・ 既存辞書の問題点として版権の問題と言語情報の欠落の問題があげられる。 ・ 言語に独立した意味情報を'computational description'の形で記述した辞書をつくることを最終目的としている。現在のところ、このような表現形式が十分でないので、シソーラスコードのかたちで記述している。すなわち、各辞書から一つのシソーラスに対してポインターが張られている構造となっている。		

III、辞書の作成について(III-3-4)

調査対象機関	(15)日本電気	(15)日本電気	(15)日本電気	(15)日本電気	(15)日本電気
管理システムの 機能概要	- 通常のエディタ機能 - 辞書内容を規則によって更新する機能(特定のデータ項目の条件付きひろいだし、条件付き更新、削除、等) - 辞書管理者による自動的管理システム	ACOSに備わったエディタ利用		- 規則によってコントロールできるスクリーンエディタ - 規則によって自動更新可能(Modify, delete など)	- 内容検索 - 条件付き更新用 rule ベース エディタ
入力方法 (データ、 エントリー、 ルール)					
問題点、意見					

III、辞書の作成について(III-3-5)

調査対象機関	(16)富士通	(16)富士通	(16)富士通	(17)東芝	(18)ブラビス
管理システムの 機能概要	辞書エディタ Fig.III-3-8参照		ATLAS-I, ATLASII 各基本辞書に準 ずる。	メンテナンスシ ステムが用意さ れているが未発 表につき割愛	Fig.III-3-9参照
入力方法 (データ、 エントリー、 ルール)	辞書作成マニユ アルはATLAS-Iのイ ンストールテー プ中に格納されて おり、ユーザのみに 提供される。		ATLAS-I, ATLAS-II 各基準辞書に準 ずる。		コンピューター と会話形式で入 力する事が可能 な為、作業 フォーマット類 は無い。
問題点、意見	基本辞書といっ ても完全に分野独立 の用語は少なく ユーザによって相 反する要求が出る 場合がある。これ をどのように調整 して「基本」の立 場を維持するかが 課題である。				<問題点> 登録すべき語数 が膨大な為、イ ンプットの簡易 化を図る必要が ある。 一インプットプ ログラムの対処 インプットされ た単語の誤り チェックを自動 的(機械的)に行 う必要がある。 一スベルチェッ ク機能等の利用

III、辞書の作成について(III-3-6)

調査対象機関	(20)日立	(22)日本IBM	(23) シストラン	(24)KDD	(25)NTT	(26)沖電気
管理システムの 機能概要		Fig.III-3-10参照		未整備である。		
入力方法 (データ、 エントリー、 ルール)		特に無し				
問題点、意見						

IV、辞書システムの運用(IV-1)

調査対象機関	(1)電総研 新明解	(3)京都大学 (長尾研) ニュー コンサイス	(4)京都大学 (長尾研) LONGMAN	(5)インタープレス (用語ファイル)	(8)東洋情報 電字林
主なユーザ	電気情報関係	研究室のみ	研究室内	冊子体の辞書作りのための 利用と機械翻訳システムの 辞書の一部として利用	科学技術、工学 系を中心に全分 野
ユーザ数	約30				約500
エンドユーザと の関係					
辞書の登録、 修正についての 体制	エンドユーザに 任せている。	なし	なし		エンドユーザに 任せている。
エンドユーザ が 修正可能な部分 と その方法	制限していな い。	なし	なし		対訳(日英ならば 英語)個々に対 しての削除のみ。 追加登録はユー ザ辞書に対して 行う。
エンドユーザ が 操作可能な部分		なし	なし		翻訳文書作成時 に検索し、必要 な語のみ印刷し 使用する。ユー ザ辞書に独自の 用語集を登録す る。
具体的な操作内 容		なし	なし		
権利関係	辞書提供時の著 作権者、著作権 との契約に従 う。	追加なし 修正なし	原データに関す る追加、修正、 改正なし	相手先との契約によって、 利用面で一部制限を設定し ている。その他は、追加・ 修正・訂正は自由としてい ます。	修正、追加され た部分のみユー ザに帰属

IV、辞書システムの運用(IV-2)

調査対象機関	(8)東洋情報 電字林 マスターの開発	(8)東洋情報 電字林 マスターの開発	(9)Mu. 科学技術庁翻訳 プロジェクト	(10)ICOT 電子辞書 システム	(11)IPA 日本語基本動詞 辞書
主なユーザ				アカデミック ユース	昭和61年4月よ り提供可能
ユーザ数					
エンドユーザと の関係					
辞書の登録、 修正についての 体制			一元管理用辞書は専門家 総合システム用はエンド ユーザ	ユーザにまかせ ている。	
エンドユーザ が 修正可能な部分 と その方法			見出し情報、形態素情 報、構文情報、意味情 報。 総合システムの編集エ ディタにより表形式編集 とLISPS式による編集 が可能。	辞書エディター で自由に修正	
エンドユーザ が 操作可能な部分			文法規則の修正は可能で あるが言語の専門的知識 を必要とするので、困 難である。		
具体的な操作内 容			TSSのディスプレイ端 末から修正が可能		
権利関係			一元管理用辞書は今後設 置される委員会等の承認 を得た上で行われる。 個人用については考え ていない。		所定の利用契約 書にもとづく。

IV、辞書システムの運用(IV-3)

調査対象機関	(14)JAPIO 特許情報機構	(15)日本電気	(15)日本電気	(15)日本電気	(15)日本電気
主なユーザ	多種	社内・商用	社内	社内	社内
ユーザ数					
エンドユーザとの関係					
辞書の登録、修正についての体制		エンドユーザと協力体制を敷いている。			
エンドユーザが修正可能な部分とその方法		エディター(修正用ツール群)Log data file for辞書修正	許可していない	エンドユーザには利用だけ開放	全ての項目に対し更新・追加可能 ・イリーガルデータの自動抽出 ・利用者入力であることを記録 ・利用者更新部は別ファイルに登録、利用時特定利用者によって利用者ファイルの優先度を高める。
エンドユーザが操作可能な部分					
具体的な操作内容	キーワード検定の場合、辞書にある用語にふられたコードを入力して論理式を作成する。				データフォームによる操作のみ
権利関係		メーカー提供分はメーカー、利用者追加・修正分は著作権に関し、半分づつ			

IV、辞書システムの運用(IV-4)

調査対象機関	(15)日本電気	(16)富士通 ATLAS-I	(16)富士通 ATLAS-II	(16)富士通 科学技術用語 辞書
主なユーザ	社内・商用	原子力、自動車 宇宙、金融		原子力、鉄鋼 宇宙、自動車
ユーザ数		70		100
エンドユーザとの関係				
辞書の登録、 修正についての 体制		エンドユーザに 任せている。	基本辞書は富士通、私用辞書 はエンドユーザと協力。	エンドユーザが 修正することはない。
エンドユーザ が 修正可能な部分 と その方法	名詞・動詞・形容 詞 辞書作成手順書 に従ってフォー マットをうめる	名詞、形容詞、 副詞、動詞、 接続詞、前置詞 感嘆詞、辞書エ ディタまたは辞 書ユティリティ による。	私用辞書 作成は科学技術用語辞書によ り分野コードでエントリを選 択、修正、追加は対話型辞書 エディタを用いるか、大量の ときはバッチ登録ユーティ リティを用いる。	
エンドユーザ が 操作可能な部分				
具体的な操作内 容		専門用語(主に名 詞)の新規登録が 大部分を占めて いる。		
権利関係	エンドユーザの 更新はNEC50% ユーザ50%	エンドユーザに 帰属する。		

IV、辞書システムの運用(IV-5)

調査対象機関	(17)東芝	(18)ブラビス	20)日立	(22)IBM
主なユーザ		医学、その他		実験段階なので、ユーザはまだいない。
ユーザ数		約300		
エンドユーザとの関係				
辞書の登録、修正についての体制	未定	エンドユーザに任せている。		未検討
エンドユーザが修正可能な部分とその方法	未定	60,000語のうち6,000語は操作できない。残り54,000語に対して削除・修正・訂正また追加が可能。辞書更新プログラムを利用。	専門用語	
エンドユーザが操作可能な部分	未定	なし		
具体的な操作内容	未定	辞書更新プログラムによる対話形式		
権利関係	未定	エンドユーザに帰属		

IV、辞書システムの運用(IV-6)

調査対象機関	(23)シストラン	(24)KDD	(25)NTT	(26)沖電気
主なユーザ		社内	なし	
ユーザ数		研究者4名		
エンドユーザとの関係				
辞書の登録、修正についての体制	エンドユーザからのフィードバックを参考意見として必要に応じて改良に役立っている。			
エンドユーザが修正可能な部分とその方法				
エンドユーザが操作可能な部分				
具体的な操作内容				
権利関係				

V、共用使用について(V-1)

調査対象機関	(1)電総研 新明解	(3)京都大学 (長尾研) ニュー コンサイス	(4)京都大学 (長尾研) LONGMAN	(5)インタープレ ス (用語ファイル)	(8)東洋情報 電子林
辞書を共用使用 するときの 問題点	フォーマット、 データの統一性	辞書そのものは検索 以外の目的に適した 構造をしていないの で機械翻訳等に直接使 用することは不可能 である。しかし、逆 に色々な工夫を考え ればシステムの目的 にあった情報を抽出 することは可能であ ろう。	辞書そのものは検索 以外の目的に適した 構造をしていないの で機械翻訳等に直接使 用することは不可能 である。しかし、逆 に色々な工夫を考え ればシステムの目的 にあった情報を抽出 することは可能であ ろう。		データ記述コー ド、 データフォー マット、 ファイルフォー マットの公開
共用使用する場 合の辞書の作成 およびメンテナ ンスで考慮すべ き点	ユーザーの組織 化による情報伝 達の効率化	本辞書の性格上他のシ ステム上の修正等が 本辞書に影響を与え ることはない。逆に 本辞書の改正は上で のべた情報の抽出の ための工夫に影響を あたえるようなもの になる可能性がある。	本辞書の性格上他のシ ステム上の修正等が 本辞書に影響を与え ることはない。逆に 本辞書の改正は上で のべた情報の抽出の ための工夫に影響を あたえるようなもの になる可能性がある。		
提供可能性	無償提供	すでに有償で外部に 提供している。		現在5点とも有償 にて発売中	有償(パッケージ 商品)
その条件	研究用 著作権者、版権 者との契約書の 交換			相手先の利用目 的により契約条 件を設定してい ます。	
媒体	MT	磁気テープ		磁気テープ	FPD(MS-DOS フォーマット)
価格		三省堂が直接有償で提 供を行っている。	英国Longman社が条 件により提供する可 能性あり。京都大学 が提供する権利はな い。	用語の場合は、 一件 40円 文章の場合は分 割不可で250～ 450万円	

V、共用使用について(V-2)

調査対象機関	(8)東洋情報 電字林	(8)東洋情報 電字林	(9)Mu. 科学技術 庁翻訳 プロジェクト	(10)ICOT 電子辞書 システム	(11)IPA 日本語基本 動詞辞書
辞書を共用使用 するときの 問題点	権利の問題	権利の問題 (共同版權である ため)	LISPのS式で表 現されている	汎用の辞書なの で、共用が困難 ということはない だろう。	利用契約書で制 限
共用使用する場 合の辞書の作成 およびメンテナ ンスで考慮すべ き点	編集要領の統一	フォーマット等 の特殊性	辞書項目に対す る修正の履歴を 各項目中に記述 する必要がある	Logical format の形式から、他 のシステムの辞 書を作成すること ができる。	
提供可能性	有償		可能と思われる	可能性あり	無償貸与 (複製の実費相当 額は利用者負担)
その条件	未検討		特定の個人、団 体が利益を得な いこと	考慮中	辞書を複製しな い。 第三者に利用さ せない。 利用状況を報告 する。 機密の保持など
媒体	未検討		MT	MT	MT
価格	未検討		未検討(低価格)	考慮中	

V、共用使用について(V-3)

調査対象機関	(14)JAPIC 特許情報機構	(15)日本電気	(15)日本電気	(15)日本電気	(15)日本電気
辞書を共用使用 するときの 問題点			意味コードの不一致 頻度データは テキストオリエン テッドであり 一般性は保障で きない。		意味コードの利用時に不一致
共用使用する場 合の辞書の作成 およびメンテナ ンスで考慮すべ き点					
提供可能性		有償	無し		
その条件		コピー厳禁			
媒体		ROM			
価格		未定			

V、共用使用について(V-4)

調査対象機関	(15)日本電気	(16)富士通 ATLAS-I	(16)富士通 ATLAS-II	(16)富士通 科学技術用辞書	(17)東芝
辞書を共用使用するときの問題点	ほぼ支障なし(意味コードに細かい不一致があるが問題なし)	ATLAS-I専用の辞書であり、共用はできない		共用はできない	見出し、数、性、人称以外はシステム固有のものであり、一般には共用できるとは思われない。 品詞、意味情報は、パーザ、意味分析器と一体のものである。
共用使用する場合の辞書の作成およびメンテナンスで考慮すべき点					パーザ、意味分析器が全く同じでなければ、共用はありえない
提供可能性	有償	システム全体は有償製品		有償	未定
その条件	コピー禁			ATLAS-IまたはATLAS-IIのユーザー	
媒体	ROM	MT		MT	
価格		35万円/月		15万円/月	

V、共用使用について(V-5)

調査対象機関	(18)プラビス	(20)日立	(22)IBM	(23)シストラン	(24)KDD
辞書を共用使用するときの問題点	権利の問題 他システムとの インターフェイス 漢字コード、 キャラクター コードの不一致		実験段階であり、まだ検討していない。	辞書の記述情報の体系がシステムアーキテクチャーによって千差万別であると思われる。その中で共通を見出し調整を計るのは非常に困難であろうと懸念される。	
共用使用する場合の辞書の作成およびメンテナンスで考慮すべき点	登録単語の共同 使用者間の合意			上記に同じ	
提供可能性	未検討			YES	
その条件	未検討			未定	
媒体	MTまたFPD			未定	
価格	未定			未定	

V、共用使用について(V-6)

調査対象機関	(25)NTT	(26)沖電気
辞書を共用使用するときの問題点		
共用使用する場合の辞書の作成およびメンテナンスで考慮すべき点		
提供可能性		なし
その条件		
媒体		
価格		

Fig. 1-1 新明解国語辞典の機能概要

- ・ F A C O M 上のシステムは、ローマ字かな変換による辞書の見出し語検索を行う。その他、品詞別語いの抽出プログラム、語釈文の分かちがちプログラムなどがある。
- ・ D E C 上のシステムは、元の辞書より変換した大規模辞書を背景にもつ自然言語処理（構文解析）システムとリンクしている。

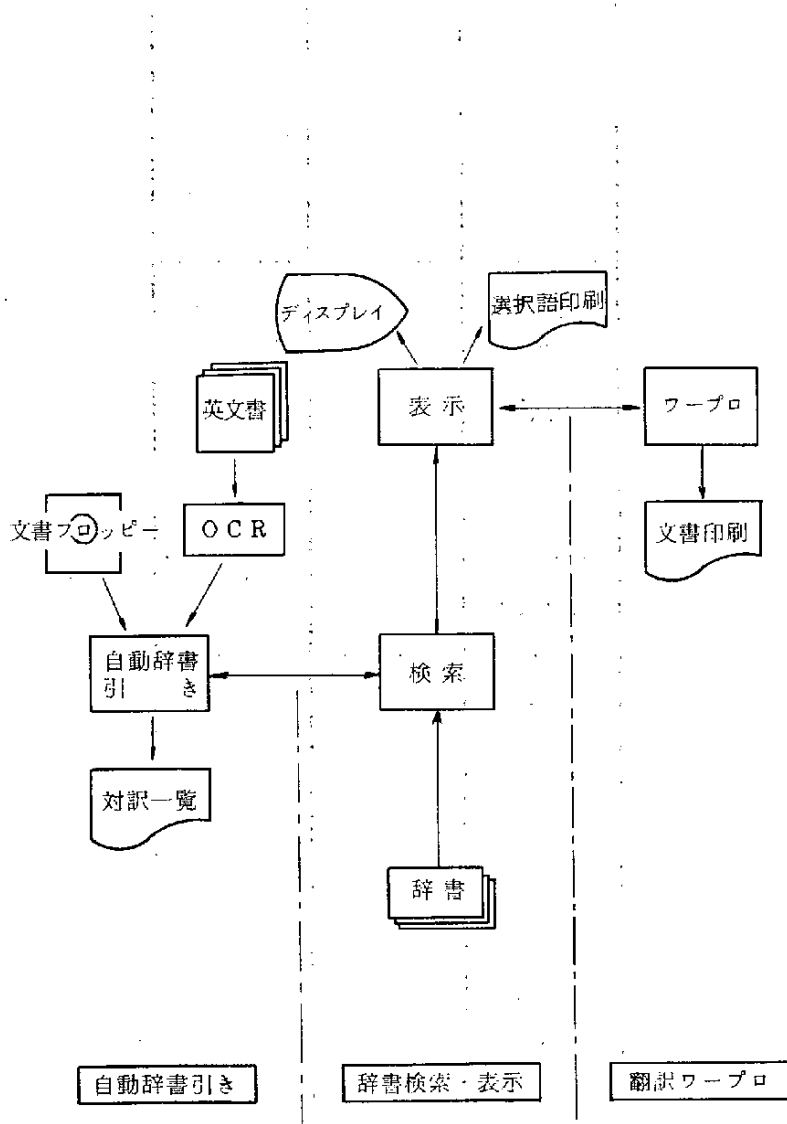
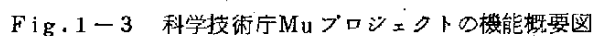


Fig. 1-2 東洋情報システム（電字林）の機能概要図



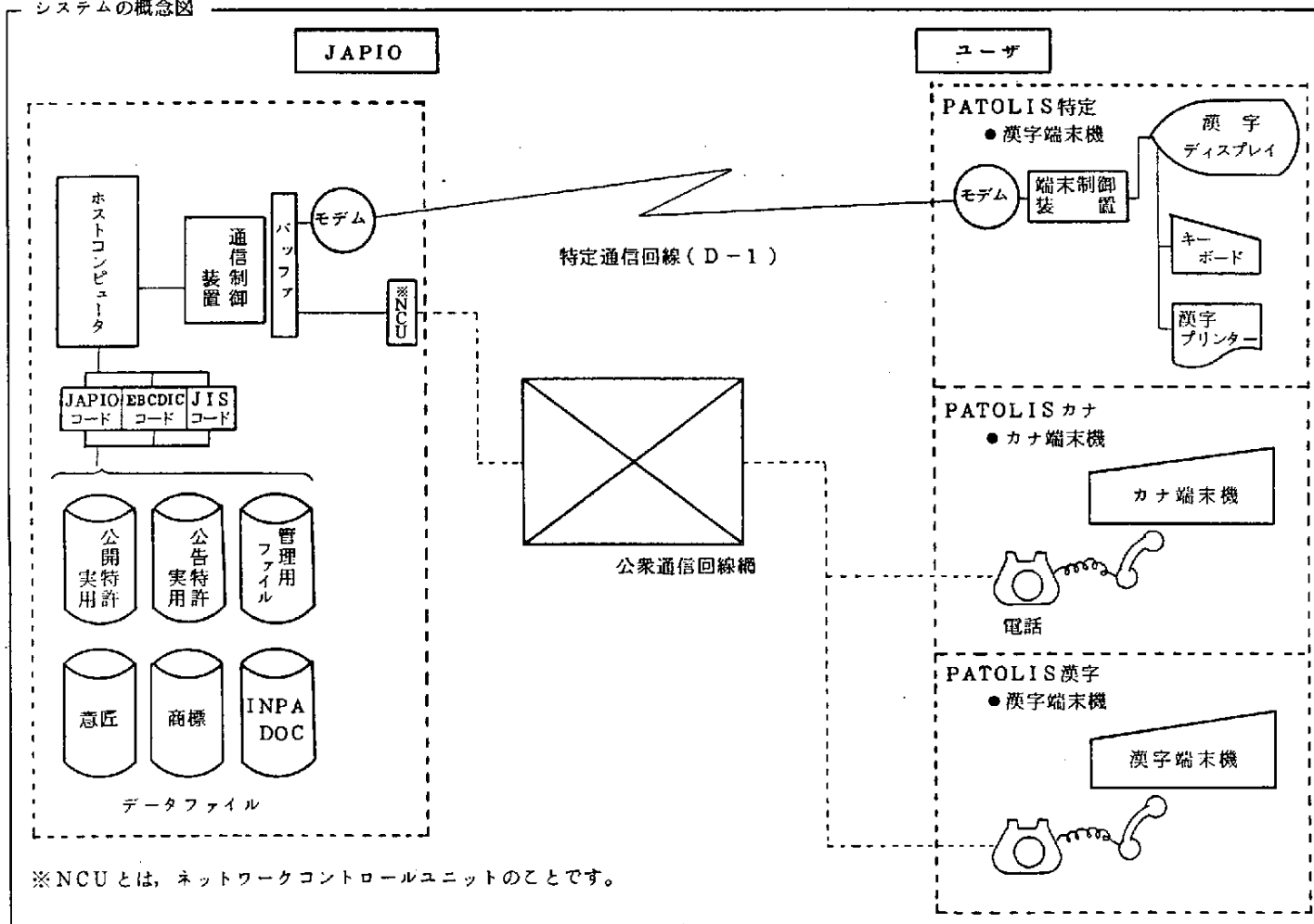


Fig. 1-4 日本特許情報機構 (JAPIO) の機能概要図

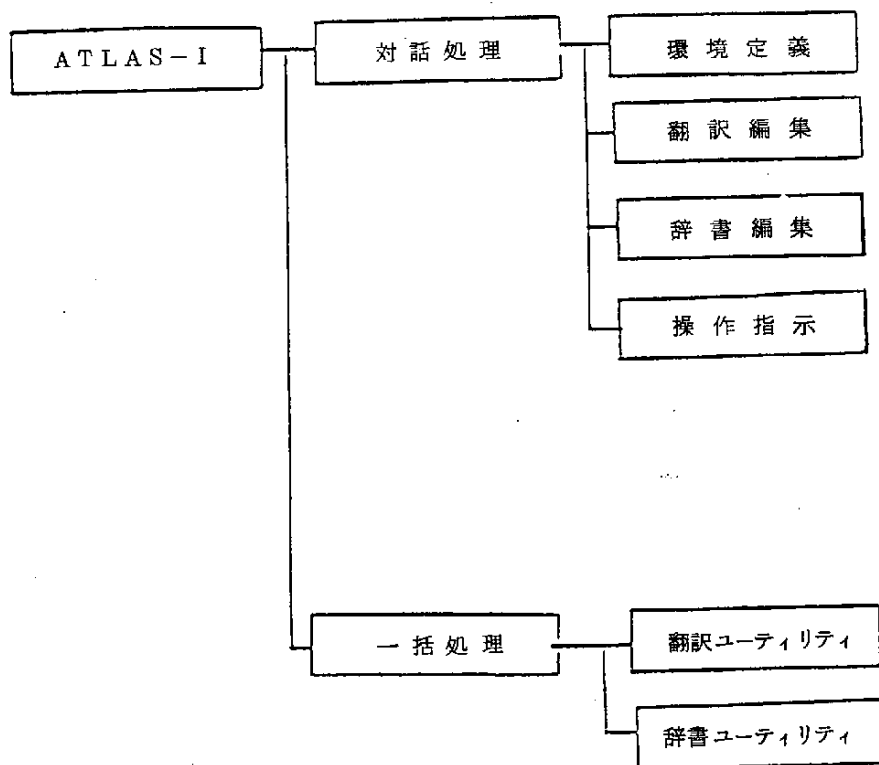


Fig. 1-5 富士通 (ATLAS-I) の機能概要図

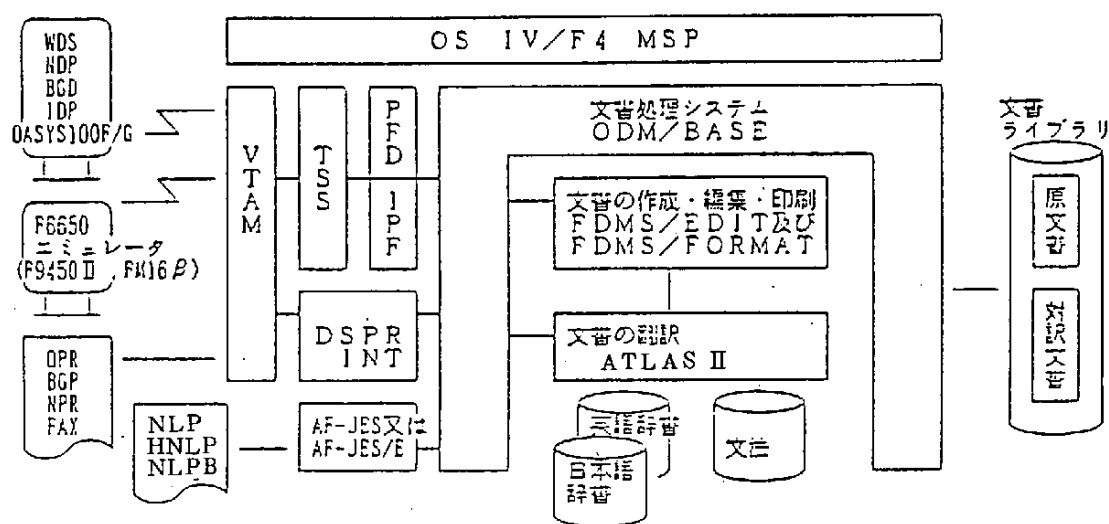


Fig. 1-6 富士通 (ATLAS-II) の機能概要図

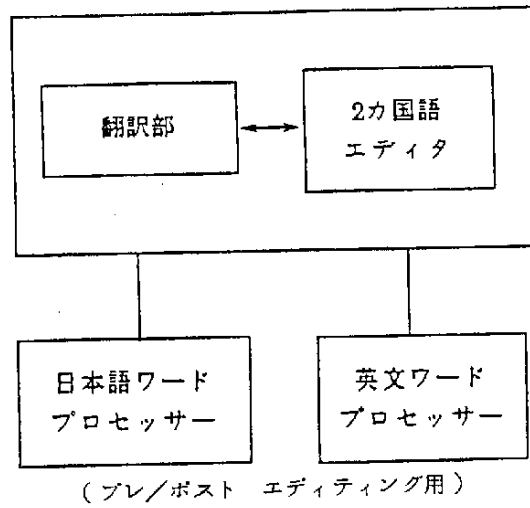


Fig. 1-7 東芝 (TAURAS) の機能概要図

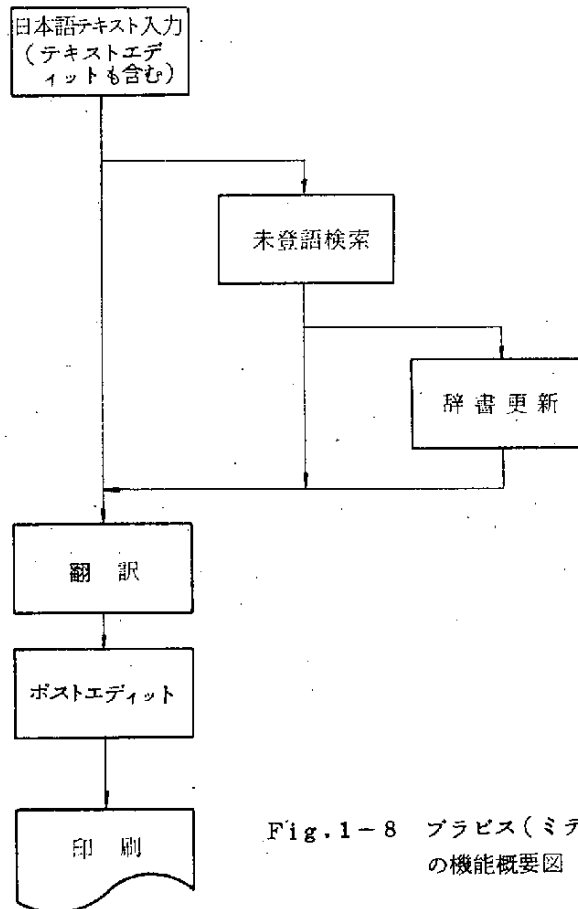


Fig. 1-8 プラビス (ミディアムパック) の機能概要図

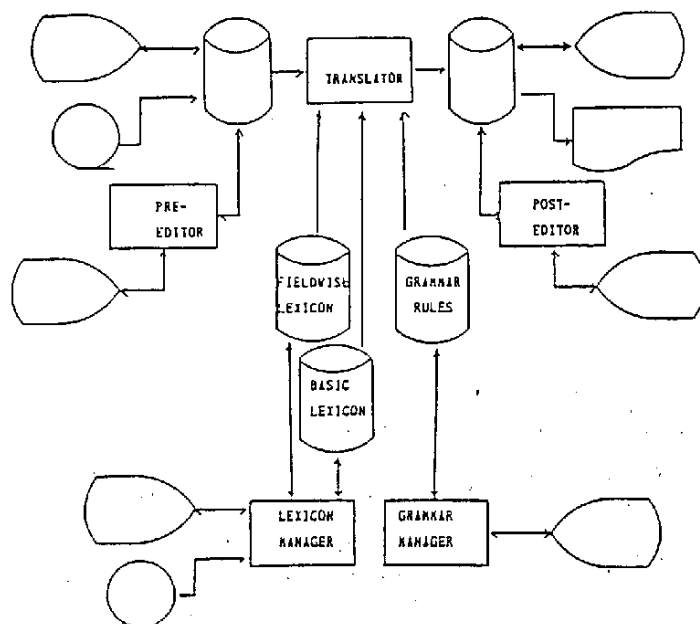


Fig. 1-9 日立 (ATHENE) の機能概要図

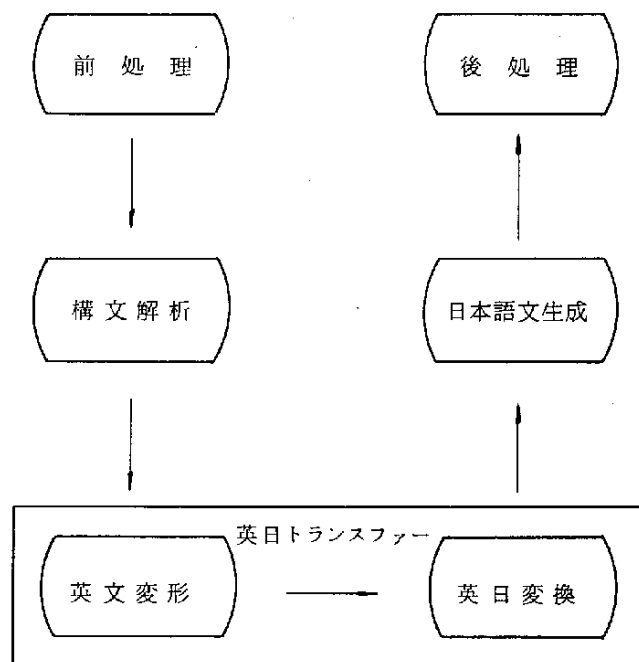


Fig. 1-10 日本IBM (SHALT) の機能概要図

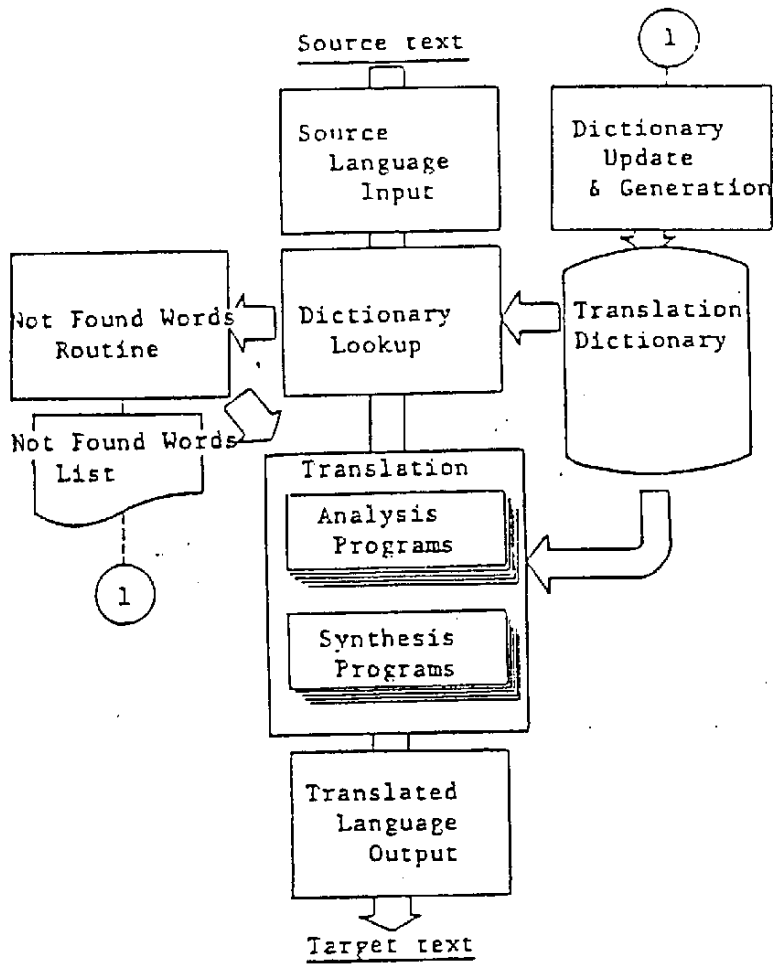


Fig. 1-11 シストラン (SYSTRAN) の機能概要図

Fig. II-1-1 東洋情報システム（電字林）の専門辞書の分野と収録語数

専門語辞書は、13分野に分かれ、目的に合わせて選択できるようになっています。提供される分野と収録語数は下表の通りです。

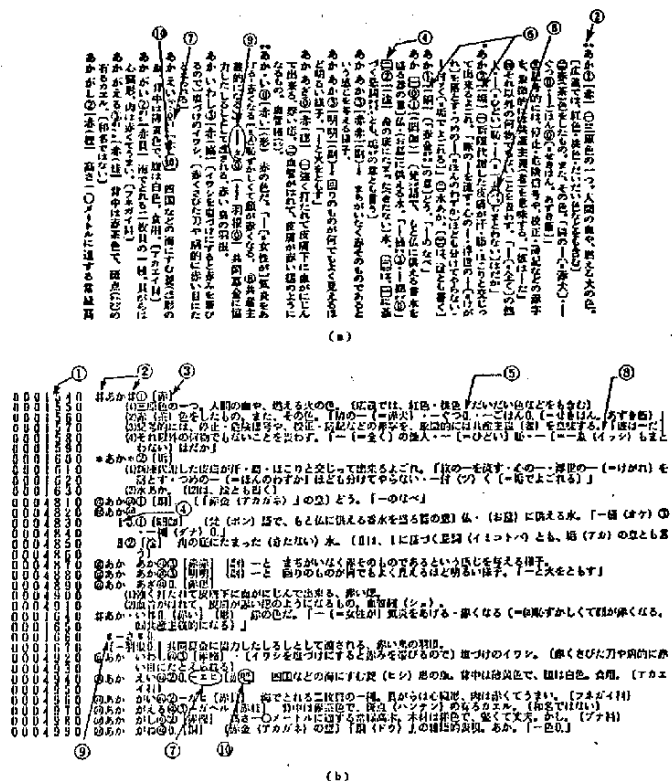
分野名	内 容	語 数
情報処理	IP・情報処理, IBM・情報処理, JIS・情報処理, IP・システムなど	30,000
電気・電子	学術・電気, JIS・電気	17,000
ビジネス	IP・ビジネスなど	80,000
プラント	IP・プラント	33,000
略 語	ANSI, MIL, IEEEなど	40,000
土木・建築	学術・土木, 学術・建築, IP・建築材料, JIS用語など	20,000
軍 事	IP・軍事	20,000
機械・設計	学術・機械, JIS・機械設計, IP・機械設計	30,000
金属・化学	学術・化学, 学術・採鉱冶金, JIS・化学・金属, IP・エネルギー, IP・公害など	27,000
サイエンス	IP・サイエンス, 学術・原子力, 学術・物理, 学術・天文, 学術・分光, 学術・遺伝など	35,000
自動車	IP・自動車, 学術・航空, 学術・船舶, JIS・自動車など	30,000
数 学	IP・数学, 学術・計測, 学術・数学, 学術・統計数学など	24,000
複 合 語	テトラ……, メタ……などの合成語を集めたもの	16,000
計		402,000

1	2.....6	9	10	11	12	13	14	64
IME	レコード data	名目	既見出し(名)……	二見出し・派生語(名)……	派生語(名)……	既見出し(名)……	二見出し・派生語(名)……	派生語(名)……

Fig. II-2-1 新明解国語辞典のデータ項目

「元の辞書と磁気テープとの違い」

- ① 原文には行番号はないが、磁気テープには JIS 漢字コードの数字で行番号（各ボリュームごとに増分 10 の通し番号、7 桁）が入っている。行番号の前後に 1 文字ずつの空白があるので、磁気テープ上の本文は、第 10 桁から始まる 55 文字で一行となっている。第 4 図では行番号が通しになっていないが、これは重要語・最重重要語は別のボリュームに入力されているためである（次項参照）。
- ② 第 10 桁目には見出し語の区分記号が来る。原文の見出し語は、ゴシックで書かれているが、磁気テープでは、前後を特殊な記号で囲むことによって区別している。元の辞書で・のついている重要語は・、**のついている最重重要語は・、その他の語は @ で前後をくくっている。また、ごく一部を除いて重要語・最重重要語は別のボリュームになっている。
- ③ 語義が複数あるものは、表記や、品詞名の後で改行し、また、各語義ごとに改行を行っている。語義は第 13 桁に番号が来る。また、元の辞書では語義を ㊦、㊧ で示しているが、磁気テープでは (1)、(2) (JEF コード 1 文字) で示している。
- ④ 読み語、同音語を示す ㊨、㊩ という記号は、ローマ数字の I、II に置きかえて、第 12 桁に示した。
- ⑤ アクセント記号は ㊪ のろは R という記号で示した。
- ⑥ 当用漢字（現在の常用漢字とは一致する）でないことを示すへ、音訓表の読み方と異なることを示すへは磁気テープの方には入力されていない。
- ⑦ 品名通いを示す片仮名小文字は、片仮名普通文字にして、同じ場所に入れてある。
- ⑧ 読み仮名を示す片仮名小文字は、前後をくくって片仮名普通文字で示した。
- ⑨ 名詞化、動詞化を示す「一さ」、「一げ」といったゴシック表記は、改行した上、前後を R でくくって 11 桁目から始まるようにしてある。小見出



第4図 (a) 元の辞書と (b) 磁気テープとの違い

- しも改行して 11 桁から始めている。
- ⑩ 前述のように、JIS コードになかったものは、見出し語の位置について、以上述べたことをまとめて第 5 図に示す。また、記号コードの対応を第 2 表にまとめて示す。
- ④中にはないが、送り仮名の任意性を示す括弧付きの平仮名（例、組（み）立（て））は、各々を丸括弧でくくった仮名（例、組（み）立（て））に展開して示している。
- §5 おわりに
- このデータは、新明解国語辞典第二版に基づくものである。その後、三省堂との間で技術指導（コンピュータ

Fig. II-2-2 新明解国語辞典のデータ項目の内容

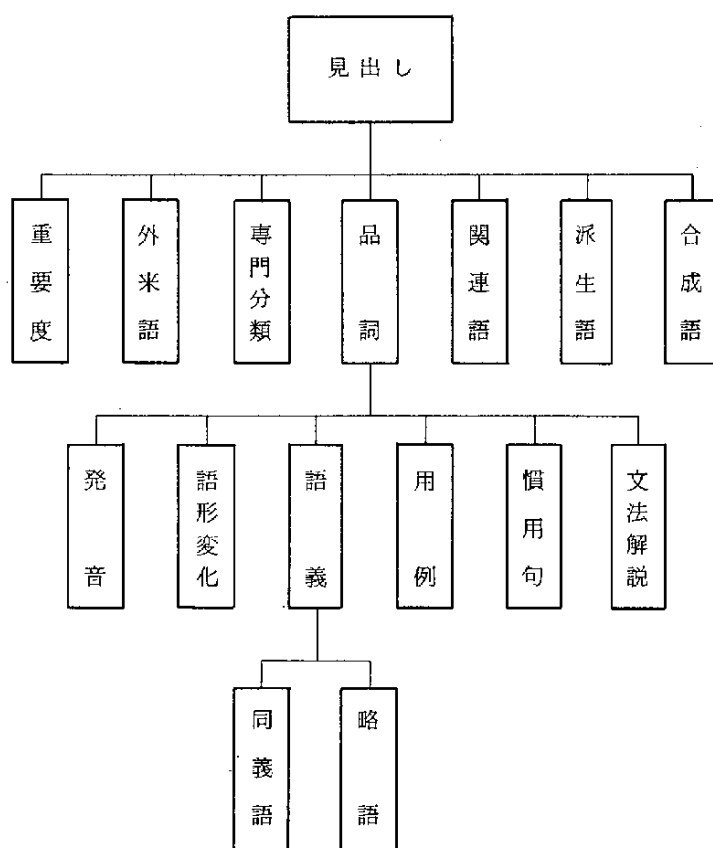


Fig. II-2-3 コンサイスのデータ項目

Fig. II-2-4 コンサイスの品詞コード一覧

コード	品 詞 名	略 号
1	形容詞	a.
2	叙述的形容詞	a.(pred)
3	限定的形容詞	a.(anrib)
4	名詞&形容詞	n., a. .注
5	副詞	ad.
6	副詞&形容詞	ad., a. .注
7	助動詞	aux.v. .注
8	名詞&自動詞&他動詞	n., vt., vi. .注
9	接続詞	conj.
10	定冠詞	def.art.
11	不定冠詞	indef.art.
12	間投詞	int.
13	名詞	n.
14	複数名詞	n., pl.
15	過去分詞	pp.
16	接頭辞	pref.
17	前置詞	prep.
18	代名詞	pron.
19	関係代名詞	rel.pron.
20	接尾辞	suf.
21	動詞	v.
22	自動詞	vi.
23	他動詞	vt.
24	自動詞&他動詞	vi., vt. .注

・注印のあるものは、略号の記述形式が一樣ではないが、同じ扱いにした。

〈例 6の副詞&形容詞は a d., a . の2とおりがあるがいずれ
a . , a d .

の場合も6とした〉

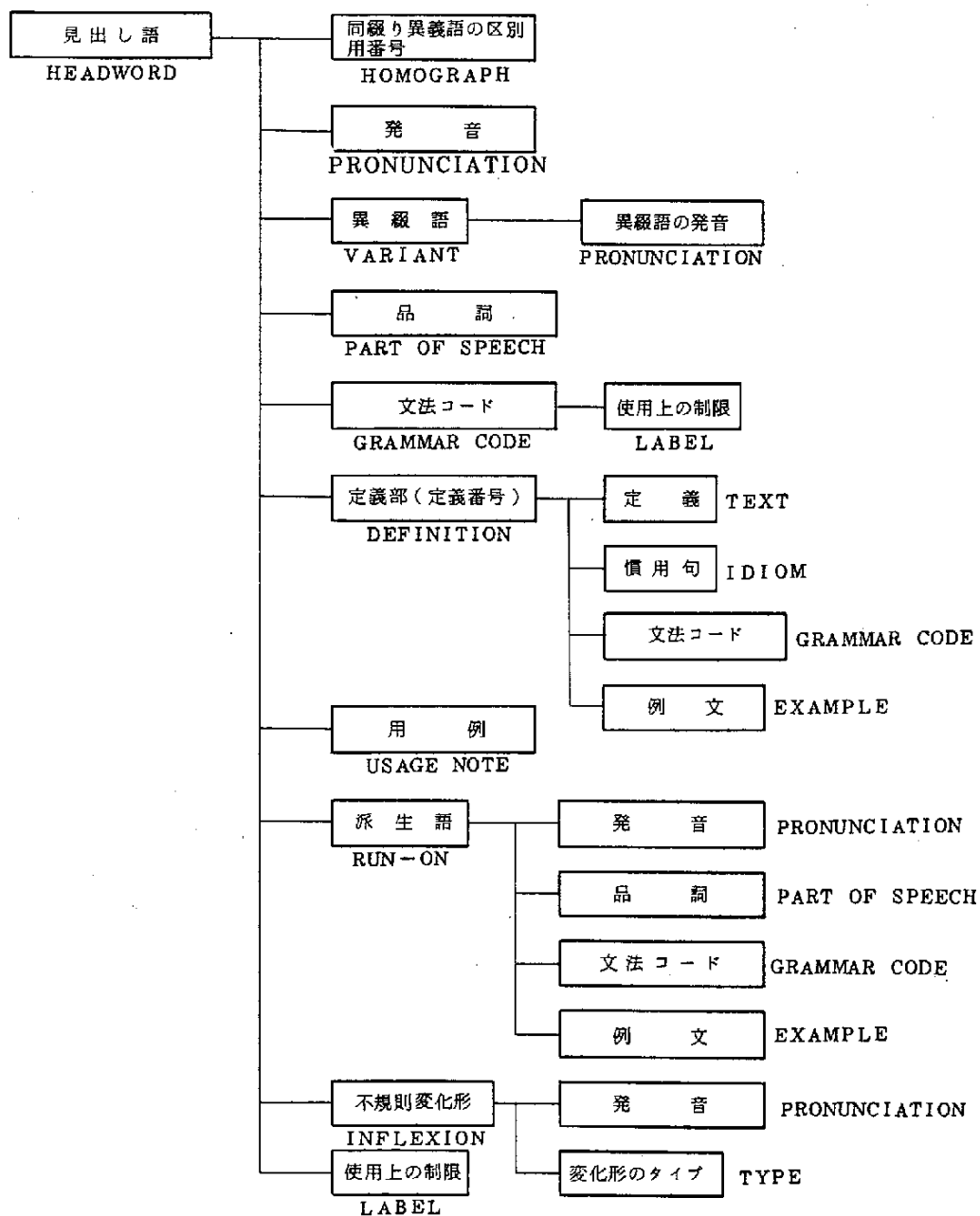


Fig. II-2-5 Longman (LDOCE) のデータ項目

大文字コード

動詞

GC	五文形	文形
[I]	1	V _i
[L]	2	V _i + C
[T]	3	V _i + O
[D]	4	V _i + IO + DO
[X]	5	V _i + O + C
[V]	5	V _i + O + INFINITIVE TO-INFINITIVE PRES. PARTICIPLE PAST PARTICIPLE

名詞

GC	
[C]	可算名詞
[U]	不可算名詞
[A]	N _i + A + NOUN
[E]	NOUN + N _i + E
[F]	複数名詞
[S]	単数名詞
[R]	i h a + N _i - R N _i - R
[N]	呼びかけ
[GC]	英合名詞

形容詞

GC	文形	
[B]	A d j + NOUN S + V + A d j S + V + O + A d j	普通形容詞
[A]	A d j + NOUN	限定用法形容詞
[F]	S + V + A d j S + V + O + A d j	叙述用法形容詞
[E]	NOUN + A d j	叙述用法形容詞
[P]	i h a + A d j	複数名詞
[GU]	i h a + A d j	単数・複数名詞

例

4. BOXES
SUGAR
GENERAL SMITH
SAADI PASHA
POLICE
A THINK
THE EARTH
GOD
GOOD-BY DOCTOR
COMMITTEE

例

BEAUTIFUL

ELDER
ASLEEP

THE PEOPLE PRESENT
THE RICH
THE ACCUSED

副詞

GC

[E]	NOUN + A d v	例
[F]	S + V + A d v	YESTERDAY
	S + V + O + A d v	
[H]	前置詞、他の副詞を伴って用いられる	RIGHT THROUGH

数字コード

[0]	V
[1]	V + NP + (NP)
[2]	V + INFINITIVE
[3]	x + i o - INFINITIVE
[4]	V + Ving
[5]	x + THAT-CLAUSE
[6]	x + WH-CLAUSE x + WH-PHASE
[7]	V + A d j
[8]	V + V ad
[9]	V + 副詞的語句 名詞、形容詞が叙述的語句を伴う

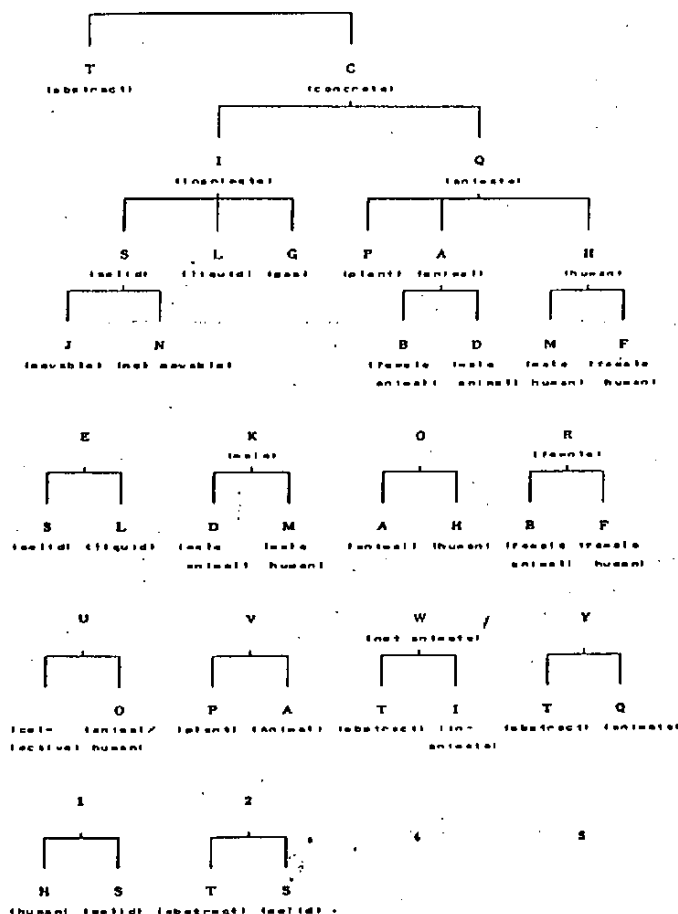
小文字コード

[1 a]	複合語	V + A d v	において名詞的表現が間に入り込むことがない。
[4 a]		V + P r e p	
[5 a]	THAT節を導く	THAT	THATが省略されてもよい。
[6 a]	WH-CLAUSE		を伴う
[1 b]	複合語	V + A d v	において名詞的表現が間に入り込んでもよい。
[4 b]		V + P r e p	
[5 b]	SO, NOTと共に用いられる		
[6 b]	WH-PHASE		を伴う
[5 c]	WISH, DESIRE 等が仮定法として	SHOULD	を伴って 用いられる。 不定詞句の中で

Fig. II-2-6 Longman (LDOCE) の品詞コード一覧

- A Animal (not human) - sex unspecified
- B Female animal
- C Concrete (not abstract) - animate or inanimate
- D Male animal
- E Solid or liquid (not gas)
- F Female human
- G Gas
- H Human - sex unspecified
- I Inanimate (not abstract) - solid, liquid or gas unspecified
- J Movable (solid)
- K Male (animal or human)
- L Liquid
- M Male human
- N Not movable
- O Animal or human - sex unspecified
- P Plant
- Q Animate (plant, animal or human)
- R Female (animate or human)
- S Solid - movable or not movable unspecified
- T Abstract
- U Collective and animal or human
- V Plant or animal (not human)
- W Not animate (abstract or inanimate)
- X Not concrete or animal (abstract or human)
- Y Abstract or animate
- Z UNMARKED (no semantic restriction)
- 1 Human or solid
- 2 Abstract or solid
- 4 Physical qualities, diseases, physical disasters
- 5 Organic material, bone, tissue, bark

(意味マーカ)



(意味マーカの階層構造)

Fig. II-2-7 Longman (LDOCE) の意味マーカ一覧

Fig. II-2-8 科学技術庁Muプロジェクト(日本語辞書)
の品詞コード一覧

構文 品詞	細分類品詞	用 例
名 (M) 詞	固有名詞 普通名詞 動作名詞(サ変) 動作名詞(その他) 副詞的名詞 格助詞的名詞 接続助詞的名詞 補文標識 未知語 式 特殊記号	(組織名, 人名, 地名等) (例)自動車, 山, 構造, 深層, 緑 「名詞+する」(例)概説, 利用, 運動 「連用形動名詞」(例)ずれ, ゆれ, ふれ (例)将来, 昨年, 従来 (例)中, 内, 前, 間, 側 (例)ため, ところ, とき, 場合, 際 (例)こと, もの (例) $f(x)$, $\exp$ (例)+, -
代(名P) 詞	疑問代名詞 人称代名詞 指示代名詞	(人, 物, 場所)(例)だれ, どれ, どこ (人)(例)彼, 彼女, わたし (物, 場所)(例)これ, あれ, それ
数(N) 詞	数 数量詞 冠数詞 助数詞	(例)1, 2, 三, 四, 百, 千, 万 (例)5 cm, 10 kg, 10 cc, 100個所 (例)第, 約, 昭和 (例)回, 件, 階, 語, 章, 部
接(Z) 辞	接頭語 接頭辞 接尾語 接尾辞	(例)諸, 各, 全, 高, 低 (例)不, 非, 反 (例)別, 上, 中, 前 (例)的, 性, 化
記 号	括弧 括弧 区切り つなぎ	(例)((例)) (例),。 (例)-/
副(F) 詞	情況副詞 程度副詞 陳述副詞 数量副詞	(文修飾, 動詞修飾) (例)結局, 極力, ここのまで (尺度, 極)(例)非常に, たいへん (假定, 疑問, 打消, 願望, 比況) (例)もし, いつ, 必ずしも (例)たくさん, 少し, おのおの
接(S) 続詞	文接続詞 句接続詞	(例)したがって, しかし, ただし (例)または, もしくは, そして

Fig. II-2-8 科学技術庁Muプロジェクト(日本語辞書)
の品詞コード一覧(つづき)

構文品詞	細分類品詞	用 例
連体詞 (R)	指示連体詞 疑問連体詞 限定連体詞 形容詞の連体詞	(例)この, その, あの (例)どの, どのような (例)ある, さる, あらゆる (例)大きな, 小さな, 少しの
動詞 (D)	動 詞	(全ての動詞)
助動詞 (A)	法助動詞 相助動詞 態助動詞 分裂構文 その他の助動詞	(否定, 必要性, 勧奨, 許可……) (相動詞を含む) (受身, 使役) (目的, 理由) (終助詞的, 補助用言的助動詞を含む)
助詞 (B)	格助詞 接続助詞 副助詞 並列助詞 終助詞 準体助詞	(例)が, に, と, で, から, より, (例)ば, と, が, のに, ので, から (例)は, も,こそ, さえ, しか, のみ (例)と, や, か (例)ね, さ, か, よ, わ (例)の, か, かどうか, か否か
形容詞 (K)	情意形容詞 性質形容詞 状態形容詞 関係形容詞	(「語幹+がる」で動詞になる) (例)うれしい, 悲しい (「もの」の性質を形容する傾向の強い語) (例)固い, 柔らかい, 新しい, 細い (「属性」や「動作」を形容する傾向の強い語) (例)著しい, 甚だしい, 早い, 強い, 高い (「こと」, 「もの」の間の関係を示す) (例)遠い, 近い
形容動詞 (L)	情意形容動詞 性質形容動詞 状態形容動詞 関係形容動詞	(「語幹+がる」で動詞になる) (例)残念だ, ゆかしいだ (「もの」の性質を形容する傾向の強い語) (例)きれいだ, 四角だ, 失礼だ (「属性」や「動作」を形容する傾向の強い語) (例)十分だ, 困難だ, 正確だ, 完全だ (「こと」, 「もの」の間の関係を示す) (例)同じ, 別

Fig. II-2-9 科学技術庁Muプロジェクト(英語辞書)の品詞コード一覧

SUB-CATEGORY OF ENGLISH PART OF SPEECH ()内は略記

Morph. P.O.S	Sub-Cat.	Comment	注1	Samples
NOUN ---(N)	PROPER (PROP)	固有名詞	○	
	COMMON (COM)	普通名詞	○	
	SYMBOL (SYMB)	記号的名詞	○	
PRONOUN (PRON)	PERSONAL (PERPRON)	case { subjective objective reflexive possessive (nominal) }	○	I, he, they, we, ... me, him, it, us, ... myself, itself, ... mine, yours, theirs, ...
	RELATIVE (RELPRON)	ditto 関係代名詞		who, which, ... that, ...
	INTERROGATIVE	疑問代名詞		who, which, ...
	DEMONSTRATIVE (DEMOPRON)	指示代名詞	○	this, that, those, these, ...
	INDEFINITE (INDEFFRON)	positive negative	○	each, all, everything, every-, many, some, some-, any, any-, either, ... none, no-, neither, ...
VERB (V)	VERB (V)	動詞	○	run, compute, catch-up, put-up-with take-care-of, take-into-account, ...
ADJECTIVE (ADJ)	ADJECTIVE (ADJ)	形容詞		Japanese, beautiful, two-legged, asleep, ...
ADVERB (ADV)	DISJUNCTIVE (DISJUNC)	文副詞(離接詞) < 話者の判断・態度等 >	○	probably, fortunately, to-be-frank, ...
	ADJUNCTIVE (ADJUNC)	(付加詞副詞) 送達副詞(動詞修飾)	○	now, seldom, before-hand, in-advance, home, yesterday, ...
	MODIFIER (MODIF)	(動詞以外の修飾)	○	very, more, quite, especially, ...
	CONJUNCTIVE (CONJUNC)	(共立的副詞) 文接続副詞	○	on-the-other-hand, therefore, however, ...
	RELATIVE (RELADV)	関係副詞		wh-, how, wh-ever, ...
	INTERROGATIVE	疑問副詞		wh-, how,
ADVERBAL PARTICLE (ADVPART)	ADVERBIAL PARTICLE (ADVPART)	副詞的小辞		off, through, up, in, ...

注1: 英語辞書で採用されている Sub-Category

Fig. II-2-9 科学技術庁 Mu プロジェクト(英語辞書)の品詞コード一覧

Morph. P.O.S	Sub-Cat.	Comment	注1	Samples
CONJUNCTION (CONJ)	COMPLEMENTIZER (COMPL)	補文接続詞		that, when, where, how, ...
	SUBORDINATER (SUBORD)	従属接続詞		because, if, as-if, rather-than, ... than(?)
	CO-ORDINATER (COORD)	并列接続詞		and, but, or, both-and, not-only-but-also, ...
CARDINAL (CARD)	CARDINAL (CARD)	数詞		one, two, ...
ORDINAL (ORD)	ORDINAL (ORD)	序数詞		first, second, ...
DETERMINER (DET)	PREDET	前置限定詞	○	all, both, double, half, twice, such, what,
	CENTDET	中置限定詞	○	this, that, these, those, ... my, his,
	POSTDET	後置限定詞	○	another, any, each, either, every, much, no, some, what, which, whose,
	POSTDET	後置限定詞	○	next, last, few, little, less, many, more other, own, same,
ARTICLE (ART)	ZERO	無定詞		-
	DEFINITE	定冠詞		the
	INDEFINITE	不定冠詞		a, an,
PREPOSITION (PREP)	PREPOSITION (PREP)	前置詞		on, of, for, to, including, based-on, in-spite-of, owing-to,
AUXILIARY-VERB (AUX)	PRIMARY	主要動詞 (それぞれ Verb としても エントリー)		do, have, be,
	MODAL	助動詞 (need, dare+2 動詞としても エントリー)		can, may, shall, will, must, ought-to, need, dare, could, should, would, might,
	SEMI-AUXILIARY (SEMI-AUX)	準助動詞		have-to, begin-to, be-apt-to, ...
UNIT		単位記号		

* CENTDET ± CENTDET, POSDET (所有代名詞), DEMDET (指示限定詞) に分解する。従つて Determiner の Sub-category は

```

├ PREDET
├ CENTDET
├ POSTDET
├ POSDET (my, his ...)
└ DEMDET (this, that ...)

```

となる。

注1: 英語辞書で採用されている Sub-Category

Fig. II-2-10 科学技術庁Muプロジェクトの(日本語辞書)格ラベル一覧

日本語名	英語名	用 例
(1) 主体	SUBject	ーが
(2) 対象	OBJect	ーを
(3) 受け手	RECIpient	ーに与える
(4) 与え手	ORIGin	ーから受ける, 奪う
(5) 相手 1	PARtner	ーと協議する, 異なる, ーに関連する
(6) 相手 2	OPPonent	ーから保護する, 独立する
(7) 時	TIME	1980年に
(8) 時・始点	Time-From	5月から
(9) 時・終点	Time-TO	来年まで
(10) 時間	DURation	5分間加熱する
(11) 場所	SPAcce	ーに位置する, ーで発生する
(12) 場所・始点	Space-From	ーから帰る
(13) 場所・終点	Space-TO	ーへ送る, ーに到達する
(14) 場所・経過	Space-THrough	ーを通る, 上空を飛ぶ
(15) 始状態	SOURce	} 5.5%から6%へ引き上げる } 英語から日本語に翻訳する
(16) 終状態	GOAL	
(17) 属性	ATTribute	適応性に富む, 欠ける, 乏しい
(18) 原因・理由	CAUSE	事故で死ぬ, ーから分かる
(19) 手段・道具	TOOL	イオン法で, ドリルで
(20) 材料	MATERial	ペーストで作る
(21) 構成要素	COMPonent	ーから成る, ーで構成する
(22) 方式	MANner	並列に, 10m/secで
(23) 条件	CONDition	焦点深度で決まる
(24) 目的	PURpose	ーに適する, 備える, 必要な
(25) 役割	ROLE	議長に選ぶ, ーとして用いる
(26) 内容規定	CONtent	ーと呼ぶ, 述べる, みなす
(27) 範囲規定	RANGE	ーについて, ーに関して
(28) 提題	TOPic	ーは, ーとは
(29) 観点	VIEWpoint	立場から, ーの点で
(30) 比較	COMpaRison	ーより大きい, ーに劣る, ーを上回る
(31) 随伴	ACompany	ーとともに, ーに伴って
(32) 度合	DEGREE	5%増加する, 3キロヤせる
(33) 陳述	PREdicative	ーである
(34) その他	ETC	

(注) 英語名中, 大文字の部分(3字)を略称とする。

Fig. II-2-11 科学技術庁Muプロジェクト(英語辞書)のグラベルー覧

(I) INNER/OUTER CASE

格 ラ ベ ル	定 義	例
(1) AGen T	Action 動詞の行為主体	• <u>The workman</u> hammered the metal flat.
(2) Causal POtency	無意志主体	• <u>This comparison</u> makes it possible to conclude that ~.
(3) EXPeriencer	知覚・感情などの精神的 事象の体験者 (Humanに限る)	• <u>The man</u> awoke to find himself in a clitch.
(4) OBJect	変化移動の対象物, 精神的 事象の対象物, 状態表 現の焦点となるもの	• She boiled <u>the rice</u> . • They agreed <u>not to oppose my plan</u> . • <u>This box</u> weighs five kilos.
(5) RECipient	所有的移動, 精神的移動 における対象物の受け手	• I showed <u>them</u> how to do it. • I showed it <u>to my parents</u> .
(6) ORIGIN	所有的移動, 精神的移動 における対象物の与え手 (AGTVになるものを除く)	• He drew his power <u>from the government</u> .
(7) SOURce	変化の始状態, 起点	• The distribution changes <u>from under cosine</u> to overcosine. • This cloth is made <u>of cotton</u> .
(8) GOAl	変化の終状態, 終点	• She will make <u>a good wife</u> . • How did they become <u>aquainted</u> ?
(9) COntent	OBJの内容	• The method seems <u>to be applicable</u> to
(10) PARTner	必須の相手 1	• The results agree <u>with the model</u> .
(11) OPPonent	必須の相手 2	• What prevents you from coming earlier ?
(12) BENeficiary	受益者	• I've bought some chocolate <u>for you</u> .
(13) ACCompaniment	随伴者	• I'll go there <u>with you</u>

Fig. II-2-11 科学技術庁Muプロジェクト(英語辞書)の格ラベル一覧

(2) INNER/OUTER CASE

格ラベル	定 義	例
(14) ROLe	役割	• We made you <u>our spokesman</u> .
(15) DEGRee	数量補語, 副詞句	• They chose me <u>as their leader</u> .
(16) MANner	様態	• This box weighs <u>five kilos</u> .
		• He carries himself <u>like a soldier</u> .
(17) RANge	範囲	• This book is <u>on semantics</u>
(18) COnpa Rison	比較	• I prefer walking <u>to climbing</u> .
(19) TOOl	道具・手段・方法	• He did it <u>by induction</u> .
(20) PURpose	目的	• We stopped <u>to have a rest</u> .
(21) Space-FRom	場所の起点	
(22) Space-AT	場所	
(23) Space-TO	場所の終点	
(24) Space-THrough	場所の経過点	
(25) Time-FRom	時の起点	
(26) Time-AT	時点・時刻	
(26) Time-TO	時の終点	
(28) DURation	時間	
(29) CAUse	原因/理由	• He died of <u>hunger</u> .
(30) CONdition	条件	• You may leave <u>by request of your employer</u> .
(31) RESult	結果	• I was so tired <u>that I couldn't go</u> .
(32) COnCessive	～にかかわらず	• She is tall <u>for her age</u> .

OP
国・機関・組織
(NATION & ORGANIZATION)

OY
生物
(LIVING THINGS)

OH 人(HUMAN-PROFESSION)
OB 動物(ANIMAL-NAME)
OP 植物(PLANT-NAME)
OX その他

OS
無生物
(NON LIVING THINGS)

ON 自然物(NATURAL SUBSTANCE)
OM 部品及び材料(PARTS & MATERIALS)
OA 生産物(ARTIFICIAL SUBSTANCE)
OI 施設(INSTITUTION)
OY その他

IO
知的抽象物
(INTELLECTUAL THINGS)

IT 理論・法則・学問
(INTELLECTUAL-THEORY, RULE)
IC 知的抽象の道具・方法(INTELLECTUAL TOOL)
IS 知的抽象の材料(INTELLECTUAL MATERIALS, SIGN)
IG 知的抽象の生産物(INTELLECTUAL GOODS)
IX その他

EO
部分
(PART)

EP 部分・要素(PART & ELEMENT)
EL 生物の器官及び構成要素
(LIVING THINGS, ORGAN, ELEMENTS)
EX その他

AO
属性
(ATTRIBUTE)

AN 属性名(ATTRIBUTE NAME)
AR 関係(RELATION)
AF 形態(FORM)
AC 状態(STATE)
AT 構成(STRUCTURE & CONSTRUCTION)
AP 特徴(NATURE & PROPERTY)
AX その他

PO
現象
(PHENOMENON)

PN 自然現象(NATURAL PHENOMENON)
PP 物理現象(PHYSICAL PHENOMENON)
PE 力・エネルギー(Power & Energy)
PB 生理的現象(PHYCOLOGICAL PHENOMENON)
PS 社会的現象(SOCIAL PHENOMENON)
PC 制度・慣習(SYSTEM & CUSTOMARY)
PX その他

SO
心持
(SENSE & FEELING)

SP 感覚・反応(Feeling & Reaction)
SI 認知・思考(Recognition & Thinking)
SX その他

DO
行動
(ACTION)

DA 行為(DOING)
DM 動き(MOVING)
DX その他

MO
測定
(MEASUREMENT)

MIN 数(NUMERIC)
MA 数値名(NUMERIC NAME)
MS 基準・標準(STANDARD)
MU 単位(UNIT)
MX その他

SA
場所・空間
(SPACE & TOPOGRAPHY)

TT
時間
(TIME)

TP 時点(TIME POINT)
TD 時間間隔(TIME DURATION)
TA 所要時間(TIME ATTRIBUTE)
TX その他

ZZ 全体のその他

この値、特定の語につける意味スロットとして以下のものがある。

WW

WT TIME
WD DURATION
WS CAUSE
WR RESULT
WM MANNER
WP PURPOSE
WA APPPOSITION
WC CONDITION
WL
WE
WV
WO
WX その他

Fig. II-2-12 科学技術庁Muプロジェクト(日本語辞書)の名詞意味マーカ体系表

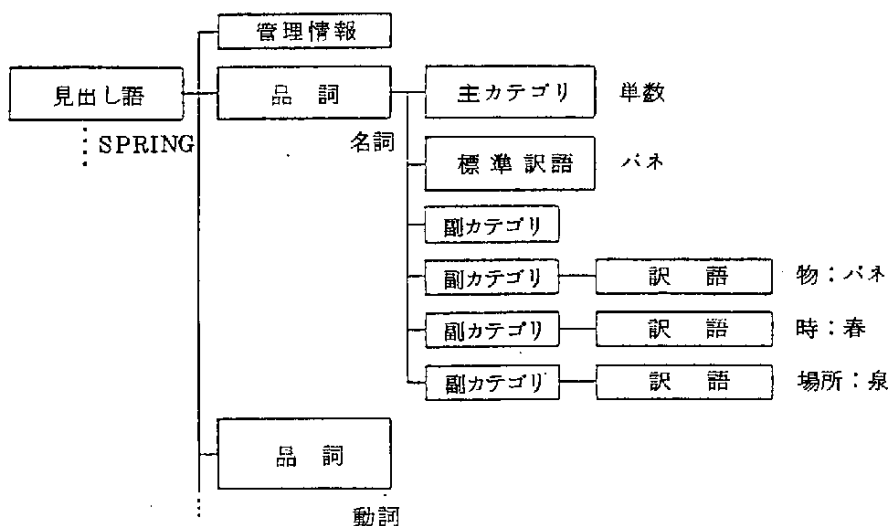


Fig. II-2-14 富士通 (ATLAS-I) のデータ項目

Fig. II-2-15 富士通 (ATLAS-II) のデータ項目

項目	データ項目名		検索
1	読み (EBCDIC)		○
2	漢字 (JEF)		○
3	概念記号 (EBCDIC)		○
4	詳細情報		
4.1	機械翻訳用情報		
4.1.1		文法属性	
4.1.2		生成規則名	
4.1.3		管理情報	
4.1.4		日付	
4.1.5		分野コード	
4.2	頻度		
4.3	隣接情報		

・述語系

動詞の記述には従来「深層格」という枠組みが採用されている例が多いが、深層格による記述には①動詞の意味記述が不十分、②執筆者ごとに記述のパラッキが大きい、という欠点を伴う。IPA Lで採用した述語系の考え方は、表層格の組み合わせのパターンと、共起する名詞句の意味素性を考慮して認定するものである。50数箇の述語系モデルを設定し、これによって統一的に記述を行なっている。

・ヴォイス

ヴォイスとは、動詞が表す動作・作用が他に及ぶのか、他からその動作・作用を受けるのか等を表す動詞の形態的变化をとらえる文法的カテゴリーである。

IPA Lではヴォイスに関し、以下の項目を記述している。

- i. その動詞が使役文（行かせる、見させる等）を形成するか否か、使役文の形はどうか。
- ii. その動詞にラレル形（行かれる、見られる等）が存在するか否か。存在するなら「直接受動」、「間接受動」、「可能」、「自発」、「尊敬」のどれを表わし得るか。
- iii. 「直接受動」を表わし得る場合の格形式の交替に関する情報。
- iv. 動詞のヴォイスの観点からの分類

・テンス・アスペクト

日本語のテンス（時制）の形式は「ル形」と「タ形」である。「ル形」とは「見る、聞く、読む」などの動詞そのままの形であり、「タ形」とは「見た、聞いた、読んだ」などの動詞の連用形に「タ」がついた形のことである。テンスとし

ての「タ形」は過去のみを表すが、「ル形」は発話時に事象が成立しているか未成立（すなわち発話時以後に成立）かによって、「現在」と「未来」を表わし得る。また「ル形」が存在しない動詞もあるので、辞書には「ル形が現在を表わす」、「ル形が未来を表わす」、「ル形が存在しない」のいずれかを記入している。

アスペクトとは、動詞の表わす動作と時間の相との関連を示す文法的カテゴリーである。例えば、「読みはじめる」は読む動作の始まることを、「読んでいる」は、読む動作が進行中であることを、「読んでしまう」は、動作の完了を表わす。

日本語のアスペクト表現はこのように「ハジメル」、「テイル」、「テシマウ」等を動詞に接続することによって得られる。IPA Lでは最も基本的なアスペクト形式「テイル形」が、進行、結果残存、経験・完了、単独状態のどれを表わすかを記述している。

また、「テイル形」の他に、「テイク、テクル、テアル、テシマウ、ハジメル、ダス、ツツアル、ツツケル、オワル」が後援可能かどうか記述している。

・ムード

ムードは、聞き手や述べる内容に対する話し手の心的態度を表わす動詞の語形変化を扱う文法的カテゴリーである。IPA Lでは以下の項目を記述している。

- i. 命令形（例：走れ、降れ）があるか、ある場合の意味は命令形か願望か
- ii. 意志形（例：走ろう、降ろう）があるか、ある場合の意味は意志・勧誘か推量か
- iii. 一タイ・タガル、一ナタイ、一ナ（禁止／願望）が後援可能かどうか
- iv. ムードによる動詞分類

Fig. II-2-13 IPA（日本語基本動詞辞書）のデータ項目の内容に関する資料

I 見出し語情報	<1>見出し語 <2>通番 <3>品詞 <4>シート番号 (i/N)	ひらがなによる表記 同音意義語を区別するための番号 最重要動詞、重要動詞の区別 N枚の下位区分の中の i 枚目を示す
II 形態情報	<5>表記 <6>異音同語 <7>活用型 <8>自他 <9>語幹 <10>派生 <11>転成	漢字と送り仮名による表記 例. つむる→つぶる 例. 五段 自動詞、他動詞の区別 語幹のローマ字表現 例. われる→わる 例. ひろがる→ひろがり
III 意味情報	<12>関連語 <13>意味記述 <14>シソーラス <15>意味分類	上位語、類義語、反義語 簡単な語釈 既存のシソーラスの分類名称、コード 意味カテゴリー
IV 構文・意味情報	<16>文型 <17>述語素 <18>意味素性 <19>名詞句 <20>文例	例. N1ガ N2ヲ N3ニ 例. LG2, O2c, A 名詞句の意味素性 名詞句の例
V ヴォイス	<21>サセ形 <22>ラレ形 <23>格形式の交替 <24>タイプ	ヲ使役、ニ使役の有無 ラレ形の有無とその意味 受身になる時の格形式の交替 ヴォイスによる動詞の分類コード
VI テンス ・アスペクト	<25>ル形 <26>テイル形 <27>その他の形式	ル形の有無とその意味 テイル形の有無とその意味 その他の形式の有無
VII ムード	<28>命令形 <29>意志形 <30>その他の形式 <31>タイプ	命令形の有無とその意味 意志形の有無とその意味 その他の形式の有無 ムードによる動詞の分類コード
VIII その他の情報	<32>慣用 <33>備考1 <34>備考2	慣用的用法 I～IIIに関して特記すべきこと IV～VIIに関して特記すべきこと

Fig. II-2-13 IPA (日本語基本動詞辞書) のデータ項目および項目の内容

名詞の副カテゴリ 2 (意味分類)

	説 明	例
-0	主張・内容・事実	FACT
-1	人	BOY
-2	動物	DOG
-3	物体	BOX
-4	量	DENSITY
-5	道具	DEVICE
-6	時間	MINUTE
-7	時刻	NOON
-8	物質	ARGON
-9	場所, 容器	ROOM
-A	行動, 動作	FLIGHT
-B	抽象物	ART
-C	物理的連動	ROT

Fig. II-2-16 富士通 (ATLAS-II) の名詞意味マーカの例

SN	MEANING
LC	LOGICAL CONTAINER
LE	LOGICAL ENTRY
LP	LOGICAL PATH
DM	DOCUMENT
ST	STATE
TH	THEORY, IDEA AND TECHNIQUE
FA	FEATURE AND ABILITY
IF	INFORMATION
AT	ATTRIBUTE
VA	VALUE OF ATTRIBUTE
HM	HUMAN
UD	UNIT OR DEVICE
WK	WORK
PS	PREDICATE/SIMPLE SENTENCE
AP	ATTRIBUTE OF PREDICATE
SL	SUPPLY FOR COMPUTER SYSTEM
PT	PART
DT	DOCUMENTATION TERMS
ML	MATERIAL
TM	TIME
PL	PLACE
PN	PERSON'S NAME
PO	POINT
OG	ORGANIZATION OR DEPARTMENT

Fig. II-2-17 日本IBM (SHALT) の意味マーカー一覧

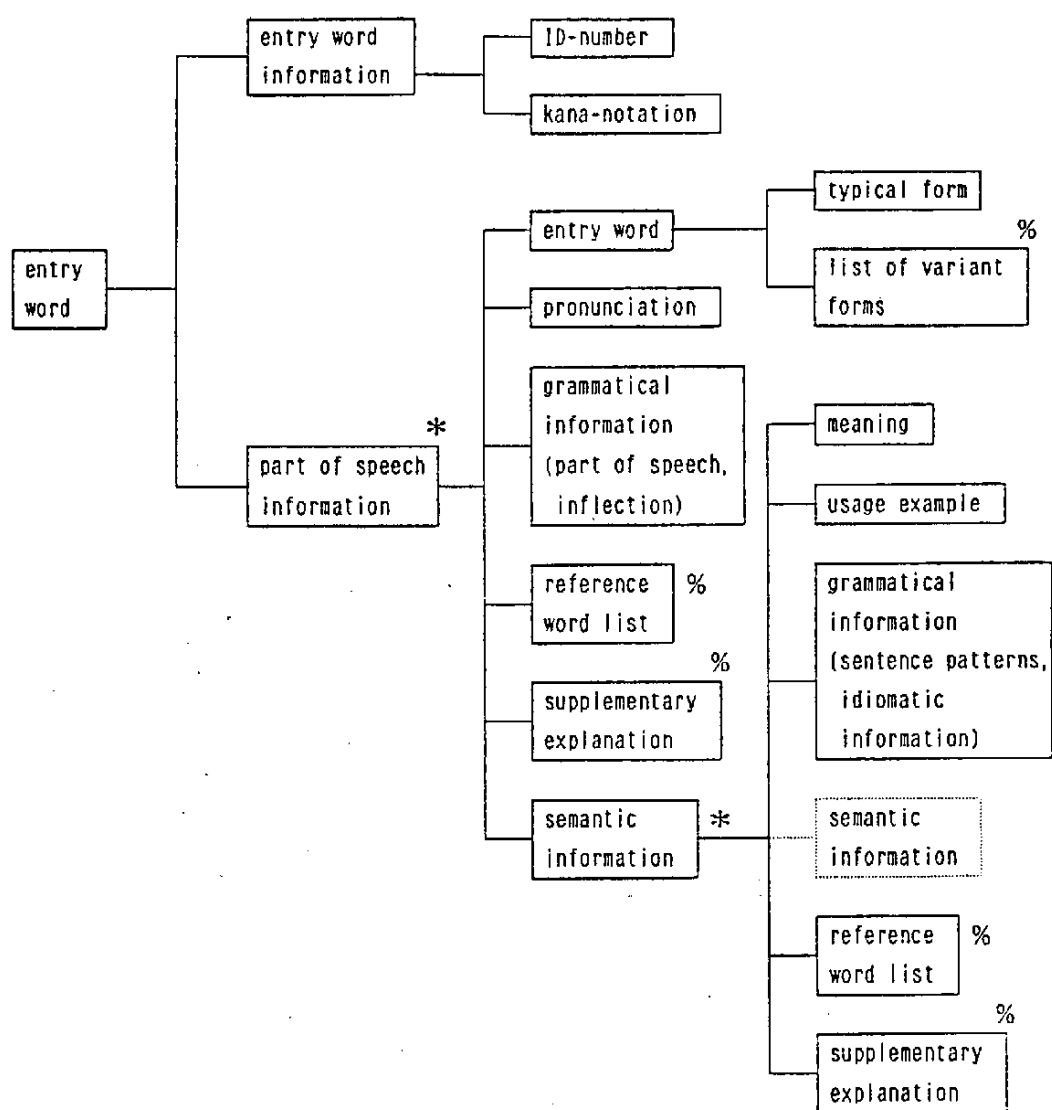


Fig. II-2-18 ICOTの日本語辞書の構造

- 165 -

1	1	1	70	20
キイ		見出しポイント	関連語	意味

93文字(186バイト)/レコード
 ブロックサイズ 1860バイト
 レコード数 18
 ファイルサイズ 13KB

関連語部

1	1	1	50	50	1
キイ		見出しポイント	派生語	意味	品詞区分

104文字(208バイト)/レコード
 ブロックサイズ 2080バイト
 レコード数 207個
 ファイルサイズ 52KB

派生語部

Fig. II-3-2 コンサイスのデータ・フォーマットの例

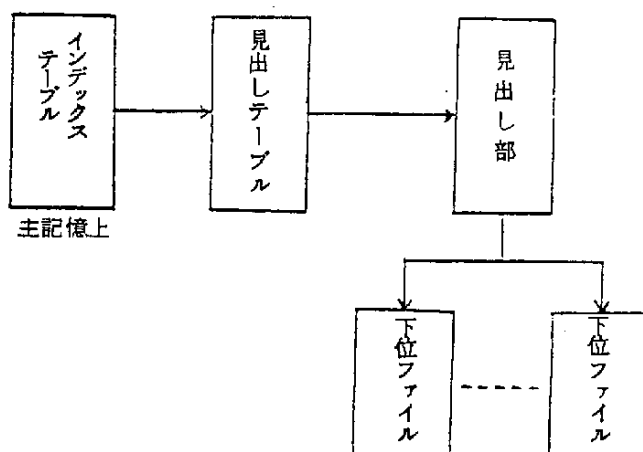
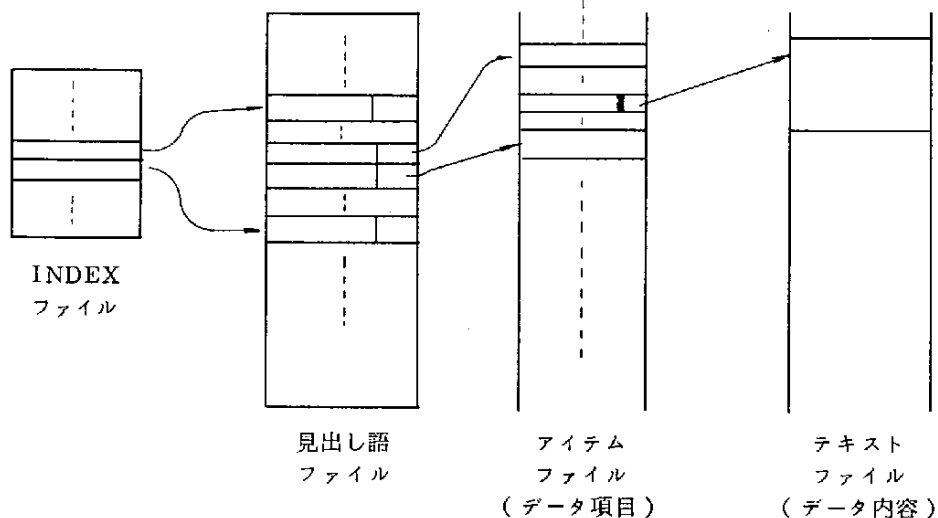


Fig. II-3-3 コンサイスのファイル・フォーマット

ISAM版



RDB版

文法コードテーブル

見出し語	品 詞	定義番号	文法コード	BOXコード
------	-----	------	-------	--------

定義文テーブル

見出し語	品 詞	定義番号	定義文 (1 定義)
------	-----	------	------------

例文テーブル

見出し語	品 詞	定義番号	例 文 (1 例文)
------	-----	------	------------

Fig. II-3-4 Longman (LDOCE) のデータ・フォーマットとファイル・フォーマット

(b) 動 詞

SEQ	E_LEX	E_CAT
		V

TRANSLATION	MAPING	E_UID	E_VERS _PAT	J_LEX		J_CAT	J_VP	J_AUX
				CASE 1	CASE 2	CASE 3		CASE 4
				E_DEEP_CASE				
				E_CON- DITION	PROPERTY			
					VALUE			
				J_DEEP_CASE				
				J_CASE_TYPE				
				J_LEX				
				J_CAT				

TRANSLATION	MAPING	E_UID	E_VERS _PAT	J_LEX		J_CAT	J_VP	J_AUX
				CASE 1	CASE 2	CASE 3		CASE 4
				E_DEEP_CASE				
				E_CON- DITION	PROPERTY			
					VALUE			
				J_DEEP_CASE				
				J_CASE_TYPE				
				J_LEX				
				J_CAT				

Fig. II-3-5 科学技術庁Muプロジェクトのデータ・フォーマットの例

```

(
  (SEQ S FIXED-NUM )
  (E_LEX S JEF-ATOM )
  (E_CAT N )
  (TRANSLATION
    ((E_UID S FIXED-NUM )
     (J_LEX S JEF-ATOM )
     (J_CAT S JEF-ATOM )
     (J_UID S FIXED-NUM )
     (COLLOCATION
       ((E_LEX S JEF-ATOM )
        (E_CAT S ESCDIC-ATOM )
        (E_UID S FIXED-NUM )
        (J_LEX S JEF-ATOM )
        (J_CAT S JEF-ATOM )
        (J_VP S ESCDIC-ATOM )
        (J_AUX S JEF-ATOM )
        (MAPPING
          (CASE1
            (E_DEEP_CASE S ESCDIC-ATOM )
            (J_DEEP_CASE S JEF-ATOM )
            (J_CASE_TYPE S ESCDIC-ATOM )
            (J_LEX S JEF-ATOM )
            (J_CAT S JEF-ATOM ))
          (CASE2
            (E_DEEP_CASE S ESCDIC-ATOM )
            (J_DEEP_CASE S JEF-ATOM )
            (J_CASE_TYPE S ESCDIC-ATOM )
            (J_LEX S JEF-ATOM )
            (J_CAT S JEF-ATOM ))
          (CASE3
            (E_DEEP_CASE S ESCDIC-ATOM )
            (J_DEEP_CASE S JEF-ATOM )
            (J_CASE_TYPE S ESCDIC-ATOM )
            (J_LEX S JEF-ATOM )
            (J_CAT S JEF-ATOM )))))
    (ASSOCIATION
      ((E_LEX S JEF-ATOM )
       (E_CAT S ESCDIC-ATOM )
       (J_DET_LEX S JEF-ATOM )
       (J_DET_CAT S JEF-ATOM )
       (J_LEX S JEF-ATOM )
       (J_CAT S JEF-ATOM )))
    (IDIOM
      ((E_PREP_LEX S JEF-ATOM )
       (E_DEEP_CASE S ESCDIC-ATOM )
       (E_N_TYPE S ESCDIC-ATOM )
       (E_ART S ESCDIC-ATOM )
       (J_LEX S JEF-ATOM )
       (J_CAT S JEF-ATOM )))
    (FOLLOWING_PREP
      ((E_PREP_LEX S JEF-ATOM )
       (E_SEM S ESCDIC-ATOM )
       (J_BXX_LEX S JEF-ATOM )))
    (PRECEDING_PREP
      ((E_PREP_LEX S JEF-ATOM )
       (J_LEX S JEF-ATOM )
       (J_CAT S JEF-ATOM )))))

```

Fig. II-3-6 科学技術庁Muプロジェクトのファイル・フォーマット

ソース語	ホモグラフ ID	訳 語	属 性	活 用 ID	クロス・ リファレンス
------	-------------	-----	-----	--------------	----------------

Fig. II-3-7 ブラビス(ミディアムパック)のデータ・フォーマット

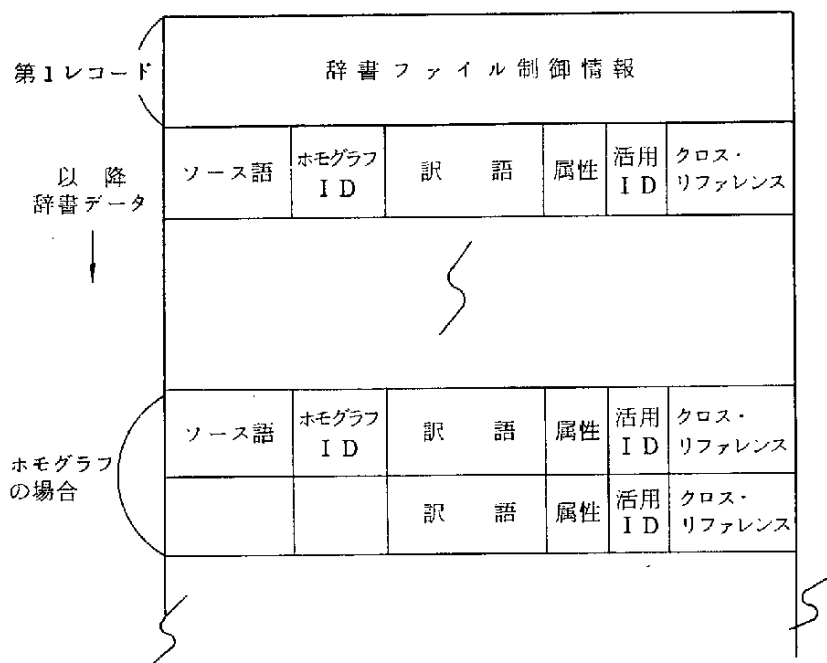


Fig. II-3-8 ブラビス(ミディアムパック)のファイル・フォーマット

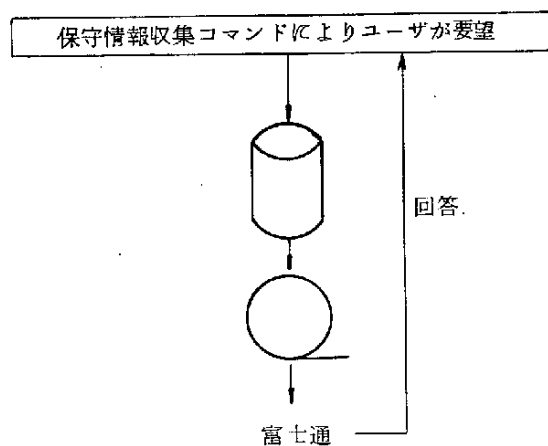


Fig. III-1-1 富士通(ATLAS-I)のメンテナンス手順

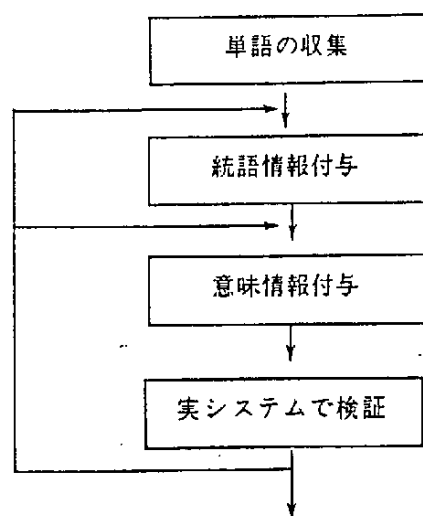


Fig. III-2-3 東芝 (TAURAS) の収録単語の分類作業

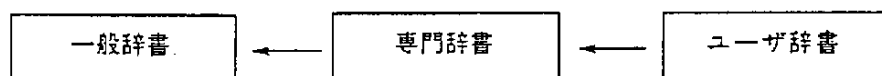


Fig. III-2-4 東芝 (TAURAS) の収録単語の選定基準

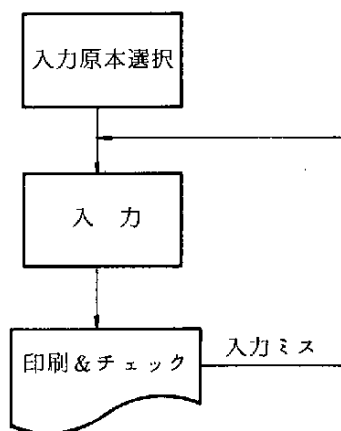


Fig. III-2-5 ブラビス (ミディアムバック) の辞書の作成手順

EDIT2 (専用エディタの概要)

コマンドの種類、記述形式はTOSBAC5600のエディタに準拠し、さらにデータの回復機能や各コマンドにカラム指定機能等を追加した。これらの追加コマンドは対象データが漢字コードであるため、できるだけコード入力の手間を省こうとしたもので、データの特性上大変便利なコマンドであった。

対象データ……1レコード32ワードのランダムファイル、高千穂コード用

プログラム……FORTRAN 約2500行

計算機環境……計算機本体はTOSBAC5600、端末はソニーエレクトロニクスの蓄像管グラフィックディスプレイ(表示速度9600ポア、1レコード漢字64文字で約1.8秒)、他に漢字鍵盤(日電のダブルレット型C5121J)を併用

Fig. III-3-1 新明解国語辞典の辞書管理システム(辞書エディター)の機能概要図

メンテナンス用システムは、現在運用していない(辞書入力時の校正用ソフトウェアは開発し、使用した)。ここには、原データから検索用データベースを作成するまでの機能概要を示す。

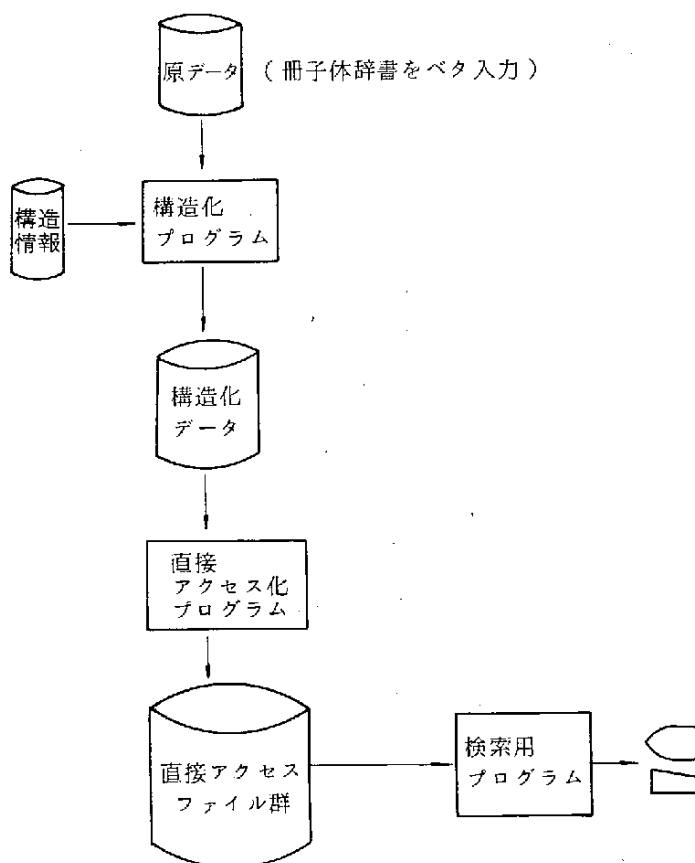
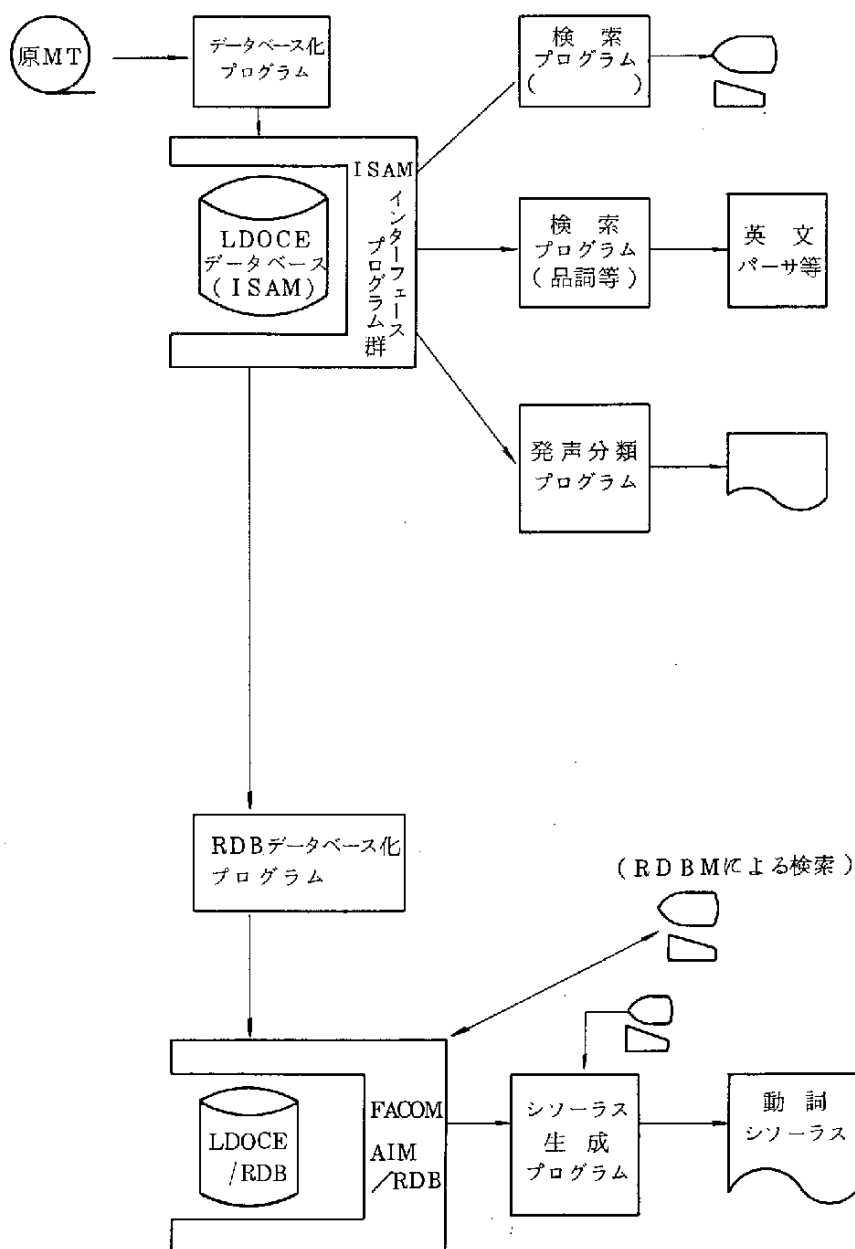


Fig. III-3-2 ニューコンサイスの辞書管理システムの機能概要図



メンテナンス用のプログラムはない。検索等利用のために京都大学で作成したプログラムについて記述する。

Fig. III-3-3 Longman (LDOCE) の辞書管理システムの機能概要図

日英，英日とも

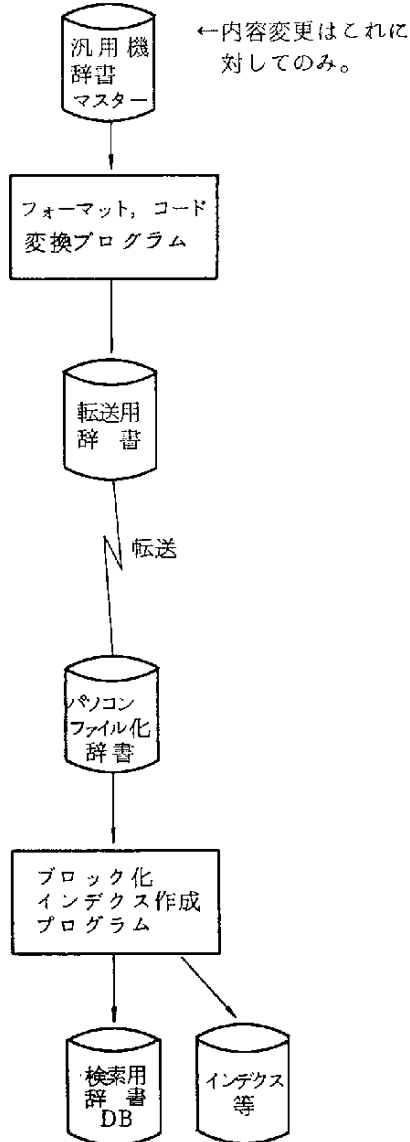


Fig. III-3-4 東洋情報システム（電字林）の辞書管理システムの機能概要図

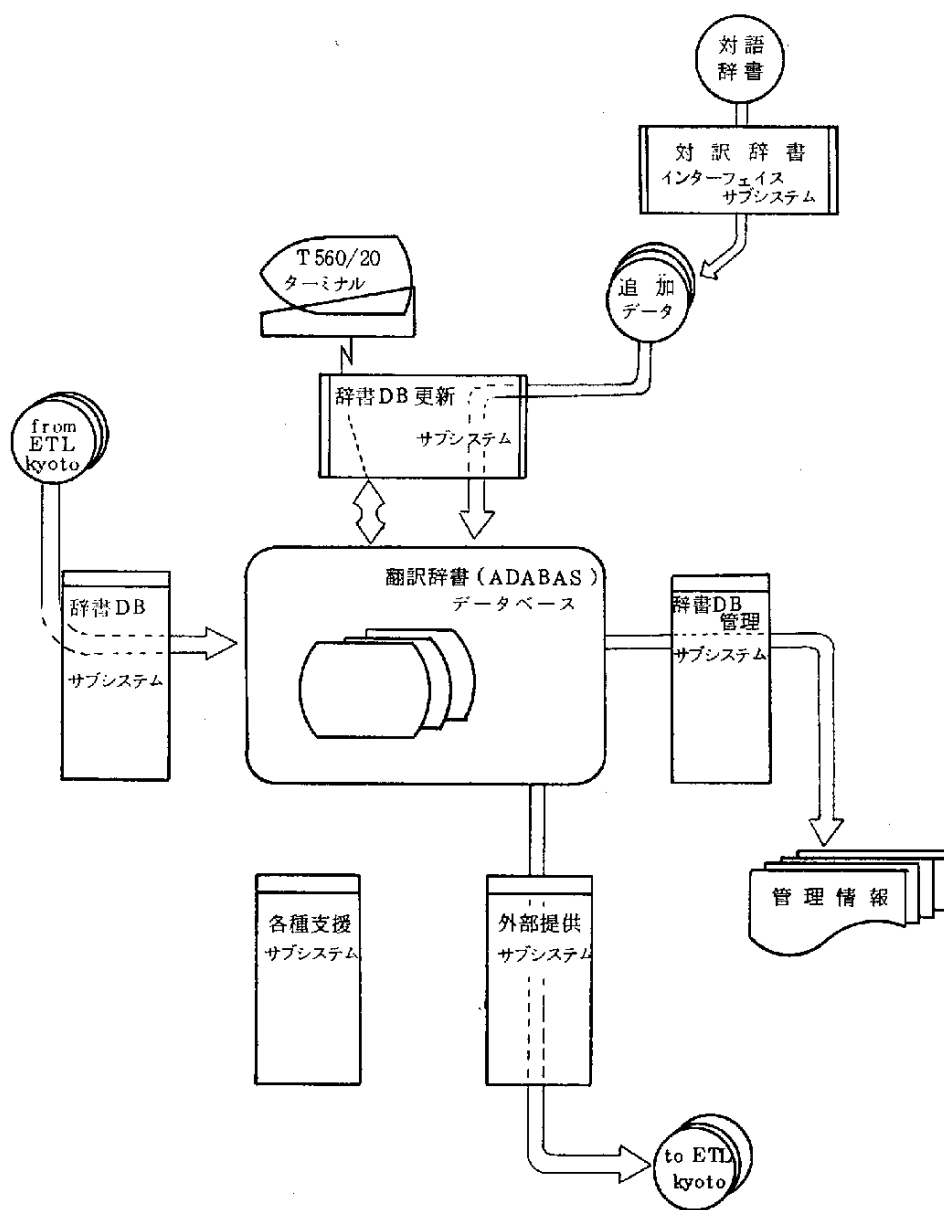


Fig. III-3-5 科学技術庁Mu プロジェクトの辞書管理システムの機能概要図

記入例

A take form → Aが具体化する

<英語動詞辞書>

form [fɔ:m] n. 1 ① 形, 形状, 形式 (shape): (1) a - 形, 形状, 外形. 2 人形, 物形. 3 型; 方式: 風切 (kind). 4 ② 形式, 外形 (= manner): 表現形式: in book - 書物の形式で, 印刷本として / in due - 正式に, 遅くも. 5 ③ 元式, 作風: It is good [bad] - to do... するの元式 [作風]だ. 6 ④ 型, 形式: (英) 表式用紙 (= 表) blank: fill in (out, up) a - 表式に記入する / after the - of... の形式と通りに / a telegraph - 電信機. 7 ⑤ (競馬) 運動家などの心身の状態, コンディション (condition): 元気 (good spirits): in (out of) - 調子が良い (悪い). 8 ⑥ (文法) 形式, 形態, 語形 (= function, meaning). 9 a) ⑦ 型, 模範例, ベンチ (教員・学生などの通例例のないもの): sit on a -, b) [英] (中等学校の) 年級 (通例 first - (1 年級) から sixth - (6 年級) である: cf. CASE 2, STANDARD 4): Tom is now in the sixth -, c) ⑧ 今 6 年生だ. 10 ⑨ (語) 形式, 形態 (= matter), as a matter of form 形式(上のこと)として, 慣習上.

for form's sake 形式を重んじるために, 形式上, in the form of... の形(型)として, の形(形式)で.
→ take form 形を成す: 目黒がつく, 具体化する.

(研究社: 新英和辞典)

..... take E_UID = 2

英大辞格	SUBJ	OBJ1	A_COL1
英小辞格	AGT	OBJ	STG
英前辞格			(O)
構文形態		(111)	
意味機能			
格必須性	1	1	1

..... take E_UID = 1

英大辞格	SUBJ	OBJ1
英小辞格	AGT	OBJ
英前辞格		
構文形態	(111)	(111)
意味機能		
格必須性	1	1

※ USAGE が複数の場合は, 深層格及び英語の場合は構文形態, 日本語の場合は意味マーカを手掛りに該当する USAGE を選ぶ.

※格必須性 1: 必須 0: 任意

COLLOCATION	E LEX			E CAT	E UID
		take		V	1
	J LEX			J CAT	J VP J AUX
		具体化 する	動詞	VZ	
	MAPPING		CASE 1	CASE 2	CASE 3
		E DEEP CASE	AGT	OBJ	
		J DEEP CASE	主格		
		J CASE TYPE		DEL	
		J LEX			
	J CAT				

<日本語動詞辞書>

..... 具体化する J_VP = V 1

日本語 主格	日本語 主格	必須
(主)	主格	1
(客)	対象	1

..... 具体化する J_VP = V 2

日本語 主格	日本語 主格	必須
(主)	主格	1

※尚, make と共起して, J-LEX が「名詞の訳語 + する」となるものは取り上げる必要はない.

※必須性 1 の格のみ転記

Fig. III-3-6 科学技術庁 Mu プロジェクトのデータ・エントリー・ルール

辞書エディタ

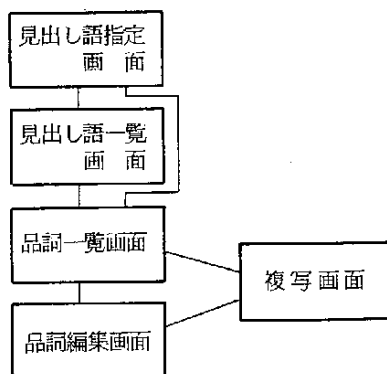


Fig. III-3-8 富士通 (ATLAS-I) の辞書管理システム (辞書エディター) の機能概要図

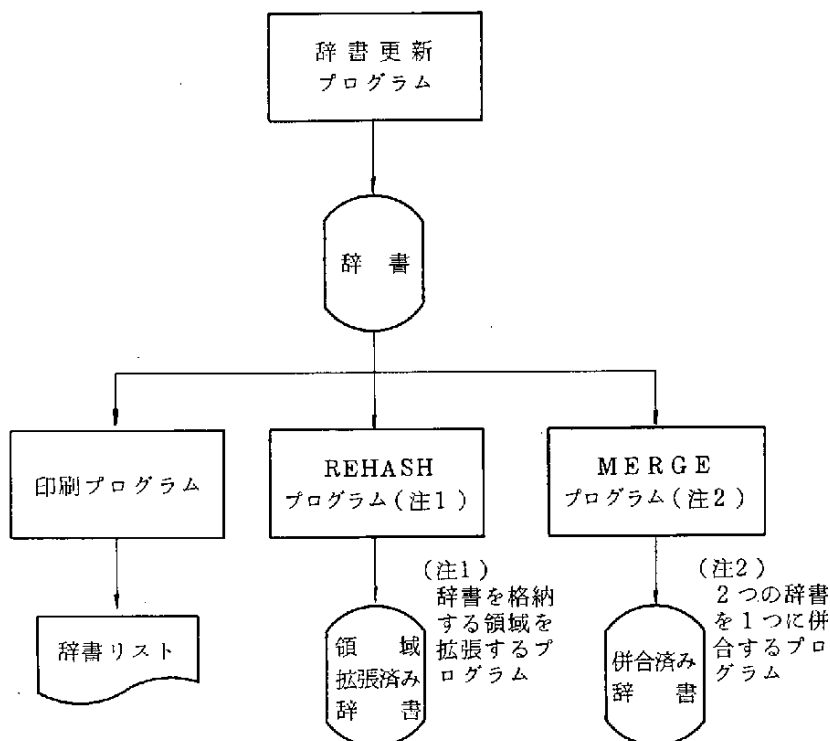


Fig. III-3-9 ブラビス (ミディアムパック) の辞書管理システムの機能概要図

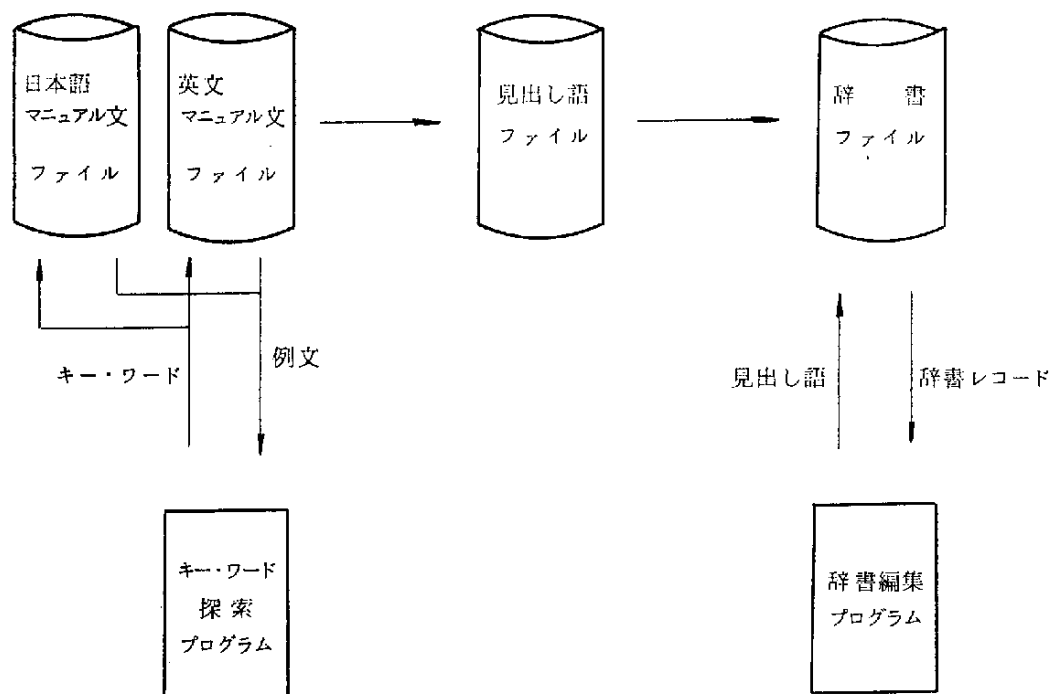


Fig. III-3-10 日本IBM (SHALT) の辞書管理システムの機能概要図

5. 自然言語解析用辞書作成の現状

ここでは共用性という観点から見た自然言語解析用辞書の作成に関わる問題点を取り上げる。

自然言語を解析する上で最も基本的なデータとしての辞書の重要性は広く認められている。さらに解析の範囲を広め、内容を深めるためにますます大規模な辞書が求められている。機械翻訳、各種自然言語インタフェースを初めとする自然言語応用システムの実用化が緒につこうとしている現時点において、辞書の共用化についての見通しを立てておくことは、将来、それらの応用システムに広く用いられる大規模な辞書を開発していく上で重要である。

以下の議論において、辞書を取りまく環境としては次のことを前提とする。即ち、自然言語の解析に用いられる理論（文法理論、意味論、…）には現在のところ決定的なものは無く、計算機上でそれを実現するアルゴリズムにしてもさまざまである、ということである。そして特にこれらを統一化の方向に持って行く必要もないし、それが可能であるとも思えない。現在の自然言語解析の研究および実用化を推進するためにこのような条件を設定することは妥当であろう。

5.1 辞書の目的および作成形態

一口に自然言語解析用辞書といっても、これにはその目的・利用形態、作成形態、規模、記述項目の種類、記述内容・方法などの面でさまざまなバリエーションがある。まず本節では、目的および作成形態の違いによる得失を考察する。

目的としては汎用のもの、特定の応用分野（たとえば機械翻訳）に専用のもの、さらには特定のシステムに組み込まれて利用されるものがある。歴史的に見れば、初期の段階の辞書はすべて特定のシステムに組み込まれたものであった。しかし自然言語処理の裾野が広がるにつれ、基礎的なデータの重複を避けるために汎用の辞書が望まれるようになってきた。

5.1.1 汎用の辞書

汎用の辞書の作成は冊子体の辞書からの変換という形を取ることが多い。現在、冊子体辞書の中にもその編集または印刷の過程で磁気テープ化されるものが増えてきた。そのような磁気テープの内容は一般には印刷形式の文字列であるので、それを計算機による効率的な検索が可能になるようにコード化かつ構造化するのである。

こうして作られる辞書には次のような得失がある。

長所

- ・大規模な機械可読辞書をきわめて短期間に開発することができる。
- ・冊子体辞書に含まれるほとんどすべての情報を整理された形で手に入れることができる。
- ・従って特定のシステムに利用したいときには必要な項目だけを抜き出して再構成すれば良い。
- ・データの信頼性が大きい。
- ・言語学（あるいは辞書学）上の問題に対する判断を、もとの冊子体辞書の執筆者・編集者に委ねることができる。

etc.

短所

- ・一方、人間の読者用の情報（語釈文、用例、語法上の注意等）が計算機には利用できない形で記述されている。
- ・目的とするシステムによっては計算機処理に必要な情報が欠落している。
- ・仮に情報が含まれている場合でも、品詞の立て方のように大雑把すぎて計算機処理には直接使えないものもある。
- ・著作権等の問題から、辞書を組み込んだ自然言語処理応用システムの商品化が難しい。

etc.

以上の点をまとめると、情報としては質が良いがそれは人間の立場から見たときの話であって、そのままの形で計算機処理に利用しようとするには無理がある、ということになる。

また、磁気テープ化されているとは言っても、それを構造化する際には次のような障害に出会うことが報告されている。

- ・誤植がある。さらに見出し語の順序に誤りがある。
- ・構造（語義のレベル等）に誤りがある。
- ・注釈等で、その対象が不分明なものがある。
- ・括弧類の使用に曖昧性がある。
- ・字種による区別（カタカナの使用等）に曖昧性がある。

etc.

これらは、仮に凡例と合っていなくとも人間ならば容易に解釈できる部分であるが、すべてを機械的に変換しようとするときの大きなネックになっている。

冊子体の辞書のデータを使う限りは常にこの問題が現われる。即ち、冊子体の辞書は人間が読むことを前提としているので、（相当厳密に体系化してあるように見えてもなお）人間の常識的な解釈にかなり依存しているのである。

5.1.2 特定のシステムに組み込まれた辞書

これは一つのシステムに必要なかつ十分な辞書的データを、そのシステムに固有のデータ構造に従って格納したものである。そこでそのシステムにとっては最適化された辞書ということができる。逆に言えば、これは本来共用化と逆行する方向を目指すものであり、あえて問題点をあげるとすれば、まさにこの点が問題となる。

即ち、他のシステムへの流用が困難である。データ構造としては、特定のシステムからのアクセスに高速に対応できるものになっているので、アクセス方法、計算機言語、あるいは計算機の機種の変更に對してきわめて弱い。また内容に関しても、特定のシステムに必要最小限のものであるので、他のシステムで要求されるものと異なるということがある。さらに文法規則を辞書に含ませているような場合、使用する文法が異なると全く使えないということにもなる。

第2点として辞書の作成あるいは拡張の際の実行体制に問題がでてくる可能性がある。なぜなら、このタイプの辞書はシステムの実験段階において小規模な形で計算機科学者によって作られることが多い。これを大規模なものに拡張するに当っては言語学者を中心とする辞書執筆者を多数動員する必要があるが、そのような場合の動員体制、執筆形式、整合性等に関して問題が生じることが考えられるからである。

5.1.3 共用化のポイント

以上の現状を考慮すると、共用化を目指す辞書の開発には次の条件を満足することが必要である。

- ・特定のシステム、機種に依存しない。
- ・汎用の形式、内容を持つものとする。
- ・計算機処理のための仕様に最初から基いて執筆する。
- ・計算機科学者と言語学者との協力体制を築く。

さらに加えれば、情報の整理、整合性のチェック等、辞書編集のあらゆる面で計算機を使うべきであるし、上記のような条件を満すことにより、それが可能になろう。

5.2 記述内容

意味論の内容および語用論の内容は次節に回すことにして、ここではそれ以外の記述項目として形態情報、音声情報、文法情報を取り上げる。

5.2.1 形態情報

変化、屈折、活用等に関しては特に問題はない。それらを見出し語として掲げるか、形態素処理と

して独立させるか、または文法的な処理の中に含ませるか等のことは個々のシステムに依存する問題であり、どの方法を取るにしても解析方法はある程度確立している。共用化辞書としてはそのためのデータを用意しておけば良いことであり、それ以上の考慮をする必要は無いだろう。

見出し語の問題に関連しては熟語、慣用句の取扱いの問題がある。これについては個々のシステムにおいても扱い方は定まっていず、また不十分であるものが多い。検索の手段も含めて検討すべき課題である。

5.2.2 音声情報

これには発音およびアクセントに関する情報が含まれると考えて良い。音声認識に使われる諸情報はむしろ物理学的な情報であり、自然言語処理の対象ではないからである。

電子化辞書の中の音声情報はむしろ合成の過程で利用されることが考えられる。現在、音声関係の基礎データの収集が進み、規則等も整備されつつあるので、今後機械可読辞書における音声情報の扱いは重要になっていく。

5.2.3 文法情報

これには品詞、格ボタン、また場合によっては文法規則等が含まれる。LFG (Lexical Functional Grammar ; 語彙機能文法) のような文法体系においては、Frege の原理により、語彙項目中に書かれた単語の意味表現が、同じくその中に書かれた単語の統語規則により他の意味表現と組み合わせられて新しい意味を構成する。しかしこのように文法に深く依存した辞書は既に述べたように共用性という目的には逆行することになる。この問題に関しては後で触れる。

いずれにしても文法情報は文法に依存する。たとえば品詞体系はそれを規定する文法抜きでは存在しない。従って、共用辞書と言えども特定のシステムで利用するときには品詞等の文法情報に多少の手直しを必要とするか、または辞書で採用している文法を用いることをシステムに要求するかということになる。これは辞書で採用している文法にどの程度の融通性があるかにもかかわる問題であり、簡単には決められない。現状では辞書の側で明確に文法を示している一部の例外を除けば、大部分は大過無しという記述方法を選んでいる。

実質的な運用効率等を考えると、辞書中に具体的な文法規則を書き下す必要が出てくることは当然である。基本的な語彙については辞書をシステムに組み込む段階で詳細な文法情報を付加することが現実的な対応であるし、また現状で多くのシステムが取っている方法である。

5.3 意味論的情報および語用論的情報

計算機による意味論の取扱い、さらには語用論の取扱いは現在実験レベルで研究が進められている

状態である。従って大規模な辞書の中に意味情報として記述するための手段も確立されていない。ここでは意味の扱いに関する例をあげるにとどめる。

機械翻訳のピボット方式のように、原文から一回その意味構造を抽出して、それから対象言語の文を生成するという方法では文の意味を扱うことが必要である。その場合、文の意味はセマンティック・ネットワークやフレームなど、ある種の知識表現形式を用いて表される。そのような意味表現は文を構成する個々の単語の意味から Frege の原理によってまとめ上げられるのである。

6. 自然言語解析用辞書の記述範囲

6.1 自然言語解析用辞書の全体像

応用システムとして英日機械翻訳を考えるならば、解析用辞書は英語に関するものになるし、日英機械翻訳ならば日本語に関するものになるが、いずれにしても理念上はそう大きな違いはないだろう。機械翻訳においては対訳辞書の類も必要になるし、解析用辞書はいつも生成用辞書と一体となって利用され、いずれか一方だけではどうしようもない。そこで、英語辞書・日本語辞書両方について考えておく。

ここでは、最初に自然言語解析用辞書の全体像をめぐっておおまかなイメージを述べたい。あとで、英語と日本語の二つに区分して、それぞれでの考察を行う予定である。

まずは、規模について考えよう。すでに作られたものなら推測するのは危険なことではあるが、それらと比較してみるならば、自然言語解析用辞書の規模（収録範囲）は数十万語レベルになるだろう。基本的な語を網羅するとともに、固有名詞や複合語を含むものを考えれば、すぐこの程度の規模になってしまうものである。現在の人間用の辞書は固有名詞や複合語がかなり弱いので、それを自然言語解析用として機械で使いこなすには、新しい見出しを大幅に追加する必要がある。そのため、場合によっては語彙調査から始めなければならない場合もある。

このような条件を考えれば、自然言語解析用の辞書として新たに作成するか、以前からあるもの（人間用であれ、機械用であれ）を転用するかのたいた問題ではなくなる。相当な実作業が必要なのである。

さて、自然言語解析用の辞書に含まれるべき情報として、どのようなものが必要だろうか。大まかな分類で示せば、次のような辞書項目を考えることができる。

ID情報：見出し・一連番号など

形態情報：表記・語形変化（動詞の活用など）

統語情報：品詞・共起するもの・連語（慣用句・イディオム）

意味情報：訳語・関連語（上位語・類義語・反対語、将来の課題として全体部分関係など）

また、文体情報も意味情報の一部として、ここに位置づけることができよう。

これらの情報をどこまで、どのように記述するかについては、個別の言語ごとの特徴があるので一概にいうことはできない。そこで、以下では英語と日本語とに分け、それぞれについてあるべき姿を考察していくことにする。

6.2 共用可能な辞書項目の範囲（英語）

ここでは、自然言語解析用の辞書として、辞書に記述される情報のうちどの範囲のものが共用可能かについて考えたい。言語としては英語を想定する。

その前に、共用ということを考える前提として、三つの問題があるので、まずそれについて簡単にみておく。

第1の問題が、応用システムごとのアルゴリズムの違いである。

現実には、応用システムごとにそこで採用されるアルゴリズムがまったくちがっており、そのため辞書に必要とされる情報も応用システムごとに違ってくるという状態である。そこで、辞書に含まれるべき情報のほとんどは、個別のシステムに必要なものなのである。ある意味で、辞書に書かれなければならない情報というのは、どういうアルゴリズムを採用するかによって決定される面が多い。

したがって、ここでとるべき道が二つに分かれる。一つは、各システムに共通に使われている情報に着目することである。しかし、このようなアプローチでは、共用辞書に記述される情報がきわめて貧弱になることが予想される。いろいろな辞書の論理積をとっていったのでは、ほとんど共通項がなくなることもあるだろう。（見出しくらいは残るだろうが。）各方面に應用可能なためには、あまり貧弱な辞書であってはいけないうけで、このへんのサジ加減が難しい。もう一つのアプローチは、論理的に「望ましい」姿（文法・アルゴリズムを含めて）について考察し、あるべき姿を目指して一から作成することである。しかし、このようなアプローチでは、作成された「共用辞書」が本当に「共用」できるのかという疑問がある。一つの文法理論にのっとって作成されれば、たしかにその文法理論を高く評価し利用している人にはつごうがいいだろうが、その他の人々にとっては何も役に立たない。結局あまたある「機械辞書」の一つにしかならないのではないか。うたい文句としては「共用」ということであっても、実態はそうではなくなるのである。

第2の問題が、文法の不完全性である。あるいは、言語の記述の不完全性といってもよい。

各応用システムのアルゴリズムの違いは、各応用システムの想定する「文法」の違いである。いろいろな文法が存在するということは、各文法が不完全だということである。では、完全な文法が作れるのだろうか。筆者は「否」であると思う。なぜならば、言語というものは、はっきり記述しようとしても記述し切れない部分が残るものだからである。

なぜ、言語を明示的にとらえることがむずかしいか。それにはいくつかの理由がある。まず、無限の事象を表現するために、有限個の要素からなる言語体系が使われるが、そのため、各要素の意味するものがいろいろあって、多義あるいは曖昧にならざるを得ない。各言語要素の意味の多義性・曖昧性が言語の本質である。現実のさまざまな意味を伝えるために言語が使われるが、まさに、この多義性・曖昧性によって人間の言語はその機能を果たしている面がある。論理的には明晰な言語のほうが望ましいとされるが、現実の言語使用においては、必ずしもそうではない。外国との交渉や政治家の発言などにおいて、多義性・曖昧性を積極的に利用する言語行動がたびたび観察される。

次に、言語は変化するということがある。変化するものがある時点で切り取って静的に記述すること自体、完全な記述ができないということである。

それから、言語の個人差である。個人語 (idiolect) という呼び名もあるくらい、実際にはいろいろな側面で言語の個人差が観察される。それを「英語」や「日本語」といった共通語でとらえるのだから、完全な記述ができないのはあたりまえである。

もう一つ、(最初の観点と似ているが)言語のある部分が外界の指示物 (referent) と関係することである。そこで、一見言語自体は変化しない場合でも、指示物が増えることによって、間接的に言語が変化したのと同じことになってしまうのである。たとえば、「黒板」の色が黒から緑色になったことを考えてみよ。そして、指示物 (外界の事物) はまさに多種多様であり、どんどん変化するものである。

このように文法が不完全である以上、(そして完全な文法はまずできない以上、)いろいろな応用システムがそれぞれの文法をそなえることは続くだろうし、それとともに利用される辞書の形態がさまざまである状態も続くだろう。

第3の問題としては、文法との関連ということがある。言語の記述が辞書と文法とでできるとするのなら、辞書はそのパートナーである文法が存在を前提としなければ、単独では記述できないといえる。文法がどんなものになるかわからないうちに、「すぐれた」辞書記述ができるだろうか。

以上のような三つの問題を考えるならば、原則として「共用辞書」という考え方は、たとえ現実の社会的要請があるとしても、それ自体あまり有効なものではない。むしろ、さまざまな辞書が(それぞれの文法なりアルゴリズムなりとともに)並立していくのが望ましいのではなかろうか。

ところで、現実に作成されている計算機用辞書が相互にきわめて類似しているとすれば、それをもとに共用辞書のイメージを描くことができるようにも考えられる。今回の国内・国外の各種の辞書の調査も、そのようなことをねらって企画されたものである。しかし、このような試みもあまりうまくいかないのではないだろうか。というのも、現存する計算機用辞書を相互に比較できるほどに詳しい情報を集めることがむずかしいからである。今回の結果についてはすでに述べられている通りだが、多くのシステムがほんの表面的なところを公開しているだけであって、辞書の実態をさらしているところはほとんどないに等しい。品詞や意味コードや、いろいろ細かい情報が付与されているようすはうかがえるが、具体的に各語にどのような品詞や意味コードが付けられているかということがわからなければ、たとえ表面的に区分が似ているからといって、辞書記述が相互に似ているとはいえないのである。

しかしながら、一方では、うたい文句としてだけでも「共用」ということを目指すことは大切なことであると思う。厳密な意味で共用ができないとしても、ある辞書に記述されている情報のうち他のシステムで比較的手軽に利用できる部分があるかもしれないからである。

辞書にとっては、どのような文法を想定するかで記述のしかたが変わってくると考えられるが、そ

れでも、各文法がかなり似た扱いをしている部分もあるので、特定の文法に拘泥するのではなく、何とかなる面もある。また、辞書の記述に際して、あまり抽象的なものは避け、言語事実に即して考えていくようにすれば、どのような文法にもある程度適用できるような辞書を作成することも不可能ではないかもしれない。

さて、一応ここではこれから実際に記述することになるのかもしれないことを念頭におきながら、一般的に共用可能な情報の範囲を自由に想像してみる。以下に述べるものの方針は、「事実」に即して記述する」ということである。なるべく明示的に記述できる項目に限定するのがよいと考えた。

(1) 見出し(表記)

これは既存のものをそのまま利用するのでよからう。分綴法くらいは当然含まれる。

(2) 語形変化

動詞の場合の時制・ing形・三単現、形容詞の場合の比較級・最上級、名詞の場合の複数形などを記述しておく。いずれも、規則的なものはルールの形で充分なので、不規則なものだけ記述すればよい。派生関係(-ly やment, -tion, un-, in-など)は、それぞれを見出しにたて、またルールでも扱えるようにしておくので、ある語のところに「派生」として記述する価値はあまりなさそうである。

(3) 品詞

従来のものを準用する。共用のためには、下位分類はあまりしないほうがよい。その理由をいくつか記しておく。

第1に、下位分類をおこなえばおこなうほど、「多品詞」語(=いくつかの品詞の特徴を兼ね備え、どれとも判別できない語)が増大し、それにともなって、記述量がふえ、記述がより困難になるという問題点がある。

第2に、下位分類の基準が曖昧であって、万人が認めるものはなかなかないということである。曖昧というよくないかもしれないが、品詞分類には形態論的・統語論的・意味論的の三つの観点があり、そのうちどれを重視するかがちがっているということである。伝統的な英文法では8品詞に分類するが、これは古典語以来の伝統であって、もともとギリシア語・ラテン語のように形態的にはっきり割り切れる言語ならあてはまっても、英語ではあまりあてはまらないものであった。そこで、Fries, Quirkなど、多くの英語学者がそれを批判し、さまざまな品詞分類や従来のものの細分類を考案してきた。しかし、結局、定説となったものはなかった。そのような大問題にあらためて立ち向かうことが果たして実り多いことなのだろうか。

第3に、下位分類をおこなうということは、「共用」を主目的とするのではなく、むしろ新しい有用な新辞書を作成するということである。そのようなアプローチをとるならば、「品詞」の下位分類などというような単純な(不徹底な)話ではすまないはずで、いろいろな「観点」をたて、それぞれについて各単語の特徴を記述するということになるだろう。将来それらを総合して(機械的に)

品詞の下位分類をおこなってもよいが、おそらく、そのような「品詞分類」はカテゴリー数が数千個くらいになってしまうのではないかと。数が大きくなるだけでなく、それは全体の見通しを極端に悪くする。すなわち、いろいろな情報を取り込んでいくということは品詞の下位分類ではできないということである。

今までのシステムでは、品詞を下位分類して、そのカテゴリーを利用して構文解析のルールを用意するのが多かった。つまり、品詞の下位分類はあくまで構文規制をより正確にするためのものだったのである。しかし、そのアプローチでいいのかどうかは議論の余地がある。下位分類がこまかくなればなるほど、ルールが複雑になる。工学的な意味で、どのくらいが最適なのかも考えておく必要がある。

(4) 共起関係

入りうる文脈と考えるとよい。入りうるかどうか、native speakerでも判断のゆれることがあるだろうが、基準としては明示的である。ごく簡単に代表的なものを述べるとすれば、次のようなものになる。

動詞の場合は、文型パターン・主語や目的語にあらわれる名詞・副詞語句・前置詞句を記述する。名詞の場合は冠詞(countableかuncountableか)・形容詞・その名詞を修飾する名詞を記述する。形容詞の場合は、それを修飾する副詞を記述する。

この情報を記述するということは、今までのやりかたでは品詞の細分類に相当することである。しかし、筆者は品詞の下位分類・細分類というとならえかたはあまりよくないと思う。品詞の下位分類・細分類という形ではなく、構文解析用の情報を「共起関係」という枠でとらえることについて、ひとこと述べておきたい。何よりもまず、それは、規則の明示性のためである。品詞分類をどうするかは一見簡単のように見えるが、すでに述べたように英語学の世界においても難問なのである。そこで、そのような難問はさておき、現実使用されている「文」について考えてみると、どのような文法体系を採用するにしても、「共起関係」はどこかであつかわざるをえない。場合によっては、それによって品詞分類を行うこともある(構文的基準による品詞分類)。そこで、大量の「文」を収集し、共起関係を分析した結果としての細かい品詞分類を示すよりは、一段前の「生の」データに近いところを示したほうが各種応用システムに利用可能になる場合が多いと考えるものである。

記述の実際についても、共起関係の記述は、品詞という形で構文的情報を示すよりもいろいろ有利な点がある。たとえば、共起関係があるかどうかについて、しばしばnative speakerの判断がゆれるものだということを述べた。もしも、共起関係をそのまま記述するとしたら、このような「ゆれ」の情報はそのまま示せばいいが、これを品詞の形で表現しようとする、いくつかの品詞間の「ゆれ」ないし、品詞の未確定というきわめて処理がむずかしい問題になる。

では、どのような範囲のものを記述すべきか。すでにその主なものは述べたが、英語の文の構造から、考えるならば、文の中心になることの多いほうから、それと共起するものという形で記述

を始めることが有効であろう。そこで、核としての動詞、さらに、名詞や形容詞などを中心にすればよいことになる。

具体的にどのような形で記述するかについては、いろいろ考慮しておかなければならない問題が山積している。単語間の共起関係という形がいいのか、あるいは、単語グループ間の共起関係がいいのかという問題にしても、確定的なことはいえない。場合によって(単語によって)特定の単語とだけ共起する場合もあるし、その結びつき方が弱くて、グループとしてとらえておいたほうがいいのか場合もあると思う。ケースバイケースであろう。

なお、このように共起関係をとらえるならば、それは純粹に構文的情報としてのみ扱うべきではない。たとえば、“～eat something”だったら、somethingのところは+edibleの名詞がくるのが普通であり、そのことは、意味的な情報(あるいは、もっと一般的に常識に近いもの)を扱っているということなのである。

共起関係を比較的多く収録している辞書としては、勝俣銓吉郎編(1958)『新英和活用大辞典第二版』(研究社)が代表的なものである。

(5) 連語

イディオムや慣用句ともいわれる。これを一つの見出しと考えて記述する場合、次のようなものが必要になろう：構文情報(全体の品詞と個々の構成要素の品詞・連語性＝結合の強さ・構成要素の置換性など)・意味情報(関連語など)・連語でない通常の意味での用法もあるか…。

共起関係という捕らえ方では充分でないような、普通の結合よりもずっと強い結合を示し、また、全体の意味がその構成部分の個々の意味の総和としては表現できないものを、ここでは連語と呼ぶことにする。ただし、実際の記述にあたっては、必ずしも両者が一義的に区分されるということではない。いつも、境界部分に位置する語があるものなのである。

連語自体は各種の辞典に記載があるので、収集することにはさほど困難はなからう。実際の記述で一番問題になるのは、連語性(結合の強さ)の判断ではなからうか。たとえば、“cool one's heels”(長く待たされる、いらいらして待つ)のような連語について、受動態(“My heels were cooled.”)が言えるのかどうか。heelsを修飾する形容語句があるのかないのか、あるとしたらそれはどういうものか。このような情報は既存の辞典類には充分記述されているとはいえない。しかし、この種の情報がなければ、構文解析すら充分にはできないであろう。

(6) 関連語

上位語・類義語・反対語、さらには将来の課題としての全体部分関係などをまとめて、関連語情報と呼ぶ。従来はシソーラスにその一部が記述されていた。

さて、意味情報については、関連語(それに訳語)だけでいいのかということは、だれもが感じる疑問であろう。しかし、もしもそれ以外の情報を盛り込むことになったら、一体どのような形で入れたらいいのか。現在のところでは、「共用」できるところまでの合意ができていないし、それ

どころか、その具体的な方法について総合的に論じているものすらないに等しいという段階である。個別の現象について、論じられることはあるし、枠組みも提示されることがあるが、実際にそれをもって、ある程度の量の記述をしたという話は聞かない。

一方、関連語というのは、一面では意味情報として不十分のように思われるかもしれないが、実はそうでもない。将来、何らかの意味記述の枠組みができたときに、それに対応して変換することもできる、応用性の広いものなのである。実際、意味のとらえかたとして、「意味は体系をなす」という考え方が有力である。これは、単語の意味はそれ単独では決定できず、そのまわりの単語群との張り合いの中で決まるものであるという考え方である。そこで、関連語についてしっかりと記述しておくことは、すべての意味情報の基本的作業となるものである。しかし、現在ではまだそこまでの意識のある人は少ないようで、関連語について記述して一体何の役に立つのかというような反応も少なくない。現在のところでは、直接の応用はあまり期待できないかもしれないが、将来の意味研究の基礎として有効である。「共用」という観点からも、誰でもある程度納得できるものを作ることはできよう。（完全なものはなかなかむずかしいが。）

今までのところでは、ツソーラスという形でまとめられているものが、ここでの関連語に近い。しかし、代表的なツソーラスである Roget's International Thesaurus にしても、単に類義語をたくさん集めてグループ化しただけあって、それらの間の関係をもっと細かくとらえるようにしないと、機械用の情報として役に立たないのではなかろうか。また、上位語の判定もおかしい。（というより、実用上の目的のために、Roget's Thesaurus では必ずしも「上位語」をグループ名にしているわけではない。）

他の関連語についても事情は同様であって、まだまだきわめて不十分な情報しかないのが現状である。これから、（人間用も含めて）新たに記述をはじめなければならない部分が多い分野である。

(7) 意味マーカ

筆者は「共用」辞書には意味マーカを含めるべきではないと考える。そこで、ここでは、なぜ意味マーカを「共用」辞書に含めるべきでないかについて個人的な意見を述べておこう。

まず、第1に、意味マーカの設定の基準がむずかしいということである。それは、意味マーカの利用方法なり目的なりがよくわからないということでもある。現在の各文法によって意味マーカの位置付けが異なるのである。また、実際の応用システムをみても、各システムで利用の仕方が決定的にちがっていると考えられるので、この情報を共用することは不可能に近いと判断する。

第2に、上に述べたこととも関連するが、意味をきちんと（明示的な形で）記述するのはたいへんむずかしいということがある。これをおおまかにいくつかのカテゴリーにまとめるのが意味マーカであるならば、極めて中途半端なものであるといわざるをえない。

現在までの研究でわかっていることであるが、意味というものは、きわめて心理的な側面があり、典型的な（prototypicalな）意味についてはおおむねの合意があっても、その周辺には曖昧な部

分が広がっていて、はっきりととらえることがむずかしい分野なのである。たとえば、Rosch などの実験によると、何が fruit かとか、何が vegetable かということを経験した人によらずに、みんなが合意するような典型的なものがあると同時に、一部の人が認めないような特殊なものもいろいろあるということである。それと同様に、apple といえば「リンゴの実」あるいは「リンゴの木」ということではほぼ全員の合意が得られても、野球のボールだとか、乳房だとかの意味があるかどうかは合意がむずかしいということである。これらは別語としての処理もできようが、apple には、必ずしもリンゴでなくともリンゴに形が似ていればそう呼べるものがあり、custard apple [パンレイシ科の植物、あるいは北米産ポーポー]、oak apple [カシワ没食子]、thorn apple [チョウセンアサガオ、あるいはサンザシの実]、May apple [ポドヒルム] など、リンゴではないものを単に apple という場合もある。つまり、apple という簡単な一語を取り上げても、いわゆるリンゴを中心にそのまわりに曖昧な領域があるわけである。

このような複雑な現象である意味を数個ないし数十個の意味マーカーでとらえることはとうていできない。(これを評して「中途半端」といったのである。) もっともっと細かいレベルでの記述をするようにところがけないと、語彙全体をとらえることはできないと思う。

英語については筆者自身があまり知らないということもあり、十分な議論ができないが、以上のようなことをもって、自然言語解析用辞書で共用可能かどうかを考える材料としたい。

6.3 共用可能な辞書項目の範囲 (日本語)

辞書の記述内容には、文法体系に依存する部分があり、それは文法理論によって異ってくる。それ故、汎用の辞書を求めるのは容易ではない。ここでは、特定の文法理論を想定しないで、特に日本語の基本的な語彙を記述するうえで、問題となりそうな項目を整理してみた。

(1) 見出し語

1. 単語の認定

- a. 形態上、単語と単語以下のレベル(形態素; 接辞, 語素)や単語以上のレベル(句)の区別があいまいなものがある。あいまいなものは、見出し語として立項しておく方がよい。漢字二文字からなる「国際」「民主」のように、見かけ上は単語相当であるが、常に合成語の要素であるという形式がある。「持っていく」「連れてくる」「きめてかかる」「かってでる」などは見かけ上は単語を越えた形式であるが、意味的なまとまり、文法的な特徴に基づくと、一単語としてみなされるべきものである。
- b. 語形変化をする単語のある語形が特殊化して独立の単語になったものも採録すべきである。たとえば、「思わず」「たえず」「残らず」が動詞の「思う」「たえる」「残る」の否定形であると同時に、独立の副詞としての用法を認めて、見出し語として立項する場合がそうである。

c. 基本語には多義語が多いが、一般に、これを一つの単語の多義とするか、複数の単語とするかの判定は困難である。「夏は暑い」「身体が熱い」「唇が厚い」の三つの「あつい」を統合するか別々の項目とするかという種類の問題がある。この際、表記の違いが意味の違いを反映していることがある。

ii. 派生語・複合語の扱い

接辞類、語素も辞書項目としてたてるとしても、「親しい→親しさ／親しみ」「女→女らしい／女みたい／女っぽい」「嬉しい→嬉しいがる／嬉しいがらせる」などの派生語をそれぞれ見出し語として立項するかどうかの基準を定めるのは困難である。

さらに複合語の問題も日本語では深刻である。特に名詞と動詞における複合語の生産性が極めて高く、複合語の全体を立項することは不可能である。複合語を構成する要素にその可能性を記述しておくことが望ましい。

接辞類には、形態音韻論の観点から異形（異音）を記述する必要がある。たとえば、真（ma- / maN- / maQ-）のように（「ま正面」「まん中」「まっ裸」）。こうした形態音韻論上の記述は複合語を構成する要素にまで及ぶと理想的である。たとえば、箱（hako / -bako）のように（「おもちゃ箱（bako）」「化粧箱（bako）」「ごみ箱（bako）」）。

iii. 表記の多様性

日本語の表記は制限がゆるく多様である。漢字とかな、それらの混ぜ書きに加えて送りがなのゆれもみられる。標準的なものと、他の可能性を区別して記述してあることが望ましい。

単語の出自を文字の種類によって表わすことも可能である。和語はひらがな、洋語はカタカナ、漢語は漢字で表記すればよい。

iv. 品詞の認定

複数の品詞にまたがるものや二つの品詞の中間的な単語群がある。以下に示す例は、いわゆる形容詞と形容動詞の形態論的な違いを見るもので、(a)は形容詞の典型例、(d)は形容動詞の典型例であるのに対して、(b)と(c)は両者の特徴が前分的に、あるいは全体として重なっているものである。

	語幹	終止	連体	連用
(a) なが		ながい	ながい	ながく
(b) ちいさ		ちいさい	ちいさい ちいさな	ちいさく
(c) やわらか		やわらかい	やわらかい やわらかな	やわらかく (やわらかに)
(d) しずか		しずかだ	しずかな	しずかに

(2) 見出し語の下位区分

一般に下位区分は、単語のもつ語彙の意味の違いによって行われる。この際、文法的な特徴の差異についても配慮されるべきである。たとえば、「ペンキを壁に塗る」と「ペンキで壁を塗る」のように、動詞「塗る」の格支配の相異は、文法的特徴の違いの一例である。

(3) 意味情報

i. 記述用語の限定

記述する側の語彙を限定することが望ましい（たとえば、基本語 2000 語の範囲で行なう）。

ii. 意味記述の方法の統一

一般に単語の意味は、意味特徴と上位語によって記述される。たとえば、「たんす」は「引き出しや開き戸があり、衣類や茶器などを整理して入れておく木製の箱状の家具」のように記述されるが、「引き出し……箱状の」の前分が意味特徴であり、「家具」が上位語に当たる。この記述を可能な限り徹底させる。外延を列挙する方法、たとえば、「金属」における「金、銀、銅、鉄、……」なども併せて利用する。

iii. 意味分類を体系化する。概念に基づくソーラスと言語よりの意味分類が考えられるが、日常語基本語の場合には意味が多様であり、分類が困難である。

iv. 百科事典的な情報

記述の範囲は、目的に応じて大小まちまちとなろう。

v. 関連語

上位語、下位語、類義語、対義語 (antonymy, converseness) などの記述。厳密に考えると少なくなるが、ゆるく多めに記述しておいて、利用の際、目的によって整理すればよい。

(4) 文法情報

(4.1) 形態に関する情報

i. 語形変化

動詞・形容詞の活用。狭義の活用のほか、接辞類のついた語形も記述の対象とする（「読む／読んだ／読もう／読んで／……」などは前者、「読みます／読ませる／読まれる／読まない／読みたい／……」などは後者）。さらに分析的な語形（「読んでいる」「読んでおく」「読んでやる」など）の可能性も記述してあると便利である。動詞「咲く」は、「咲いておく」「咲いてやる」といった分析的な語形をとらない。語形変化のタイプを示してあることが望ましい。伝統的な国文法の活用の分類だけでは情報が少なすぎる。名詞の格助詞のつき方も、一様ではない。「さつき」「先刻」が名詞であるとしても、「から」「まで」「の」は後置するが、「が」「を」「に」などがこれらの単語につくことはない。

ii. 形態論的なカテゴリー

形態論的なカテゴリーとは、ある語彙の意味をもつ単語の、少なくとも二つ以上の文法的な意味や

機能の点で対立する語形の系列からとりだされたものである。日本語の動詞には、ヴォイス、アスペリト、肯定否定、スタイル、テンス、ムードなどのカテゴリーが認められる。これらのカテゴリーの中にはさらに、いくつかのサブカテゴリーからなるものもある（たとえばヴォイス）。こうした形態論的カテゴリーについては従来の辞書には極めて不十分にしか記述されていない。IPAの「基本動詞辞書」が、こうしたカテゴリーをある程度まで考慮していて参考になる。

(4.2) 統語に関する情報

i. 支配語による語結合 (collocation)

動詞の格支配のパターンがその例。支配語になれるのは、動詞のほか、形容詞、名詞、副詞の一部である。ちなみに、陳述副詞、連体詞、接続詞、感動詞の単語群は支配語の資格をもたない。

なお、格支配については、必須要素と任意要素の問題があり、この区別は判定がゆれる。任意的な格成分を辞書の中に統一的に記述する方法は今後検討されなければならない。

また、この格成分に意味的な関係をかぶせた、いわゆる深層格は不安定な性格をもっている。安定した枠組みを得るのは容易でない。

ii. 形式語の記述

一般に形式名詞や形式動詞と呼ばれている語については、特に統語的な情報を詳細に記述しなければならない。「こと」「よう」「の」などの名詞性の形式、「する」「ある」「なる」などの動詞がその具体例である。こうした形式語の研究が近年盛んに行われており、辞書にもその成果がとりこまれなければならない。

(5) 文体 (広義) 情報

i. 歴史的

新語、古語、流行語などの区別。

ii. 地域的

方言差 (英語の場合なら、米語かイギリス語かの区別も)。

iii. 社会的

政治、経済、スポーツなどの分野ごとの差。学問領域による専門語もここに属すると考えられる。学生用語、幼児語などの区別もふくまれる。

iv. 機敏的 (あるいは、狭義の文体)

中立的な語に対して、あらたまった用語やくだけた用語 (俗語) の区別など。一般に、一つの文や文章の中で、異なる機敏の特徴をもつ用語は共起しないものである。「めし (くだけ) をめしあがる (あらたまり)」 「あなたさま (あらたまり) の小倅 (くだけ)」といった語結合がよくないのはそうした個々の単語のもつ機敏的特徴がシンタグマの中で一定していないことによる。

(6) 慣用表現

i. 強い慣用表現

「気がつく」「道草をくう」「役にたつ」「あぶない橋を渡る」などの慣用句。これらの慣用句は、(a)構成要素の結びつきが不規則で、特殊である、(b)全体の意味を要素に分解できない、(c)形態上かつ統語上の拘束が強くはたらく、(d)あらかじめ形づくられた既製品である、(e)全体で単語相当である、など。特徴をもっていて、慣用句そのものが辞書の見出し語として登録されなければならないものである。

ii. 弱い慣用表現

「ブレーキをかける」「ブラシをかける」「アイロンをかける」「スイッチを入れる」「はさみを入れる」などの表現。こうした慣用表現は辞書の中で例示しておくことが望ましい。「ブレーキをかける」の例なら、「ブレーキ」の項にも「かける」の項にも、この例示がほしい。ちなみに、このようなタイプの慣用表現は、言語によっては一単語で表わされるものが比較的多い。

7. 自然言語解析用辞書の運用

7.1 辞書の共用

はじめに、自然言語解析用辞書の共用について辞書の運用の立場で考察してみる。

言葉を扱うシステムであれば、大なり小なりあるいはアルゴリズムの複雑さの度合は異なっても何等かの形で自然言語解析用辞書の利用が不可欠である。しかし、これから新しく研究なり開発を始めようとする、まずそこで使用する辞書を用意しなければならない。辞書を始めから作るのは大変なので一般に冊子体辞書などを基本にして開発が進められる場合が多いようであるが、開発期間・費用などの問題が大きい。このような時に機械可読な既存の辞書が利用可能であれば、それらの適当な辞書から必要な素材を抽出して加工したものを基本に組み合わせて、不足の情報を付加して新しい辞書を作成する方法が考えられる。また、既存辞書から不足している情報を見つけ、それを素材にして自らの辞書に付加的に取り込む方法も考えられる。

このような辞書相互の共用を目的とした“共用辞書フォーマット”の検討も必要とされよう。共用辞書フォーマットが実現可能であれば図7-1に示すように共用辞書フォーマットを経由した変換方式を採用することによって、共用を効率的に促進することが可能となり、かつ、運用時の辞書の改定に当たっても相互の共用の可能性があると思われる。

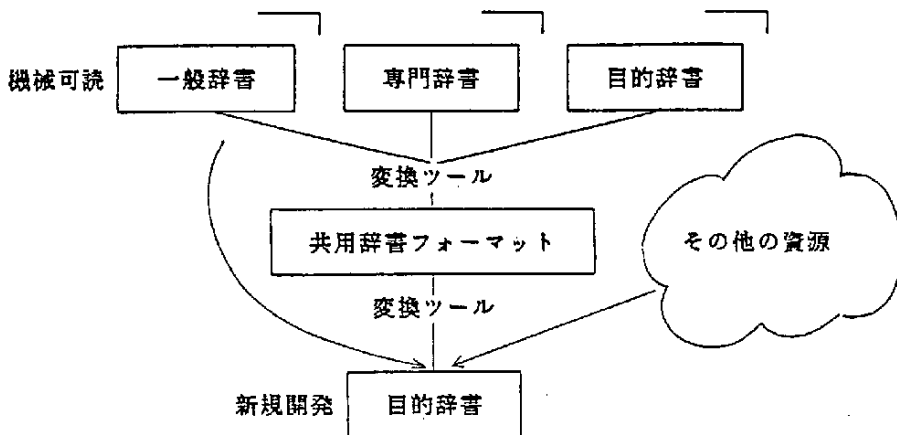


図7-1

また、一旦辞書の開発が終了してもその適用システムのニーズに対応して絶えず手直し、改良・拡張がせまられることになる。このためには開発の体制に引き続き運用体制への移行をきちんとしておく必要がある。一方ユーザにとっては、辞書利用のし易さと共に維持管理の容易性も重要である。

7.2 辞書運用上の考慮点

次に、ユーザが自然言語解析用辞書を応用システムに運用していく上での幾つかの考慮すべき点について考察してみる。応用システムとしての機械翻訳、自動インデックス、OAやデータベース検索でのマンマシンインタフェースなどに対応して辞書の機能レベルは様々であり運用形態もそれぞれ異なると思われるが、ここではそれらを個々に論ずることなく一般論として述べることにする。

自然言語解析用辞書をユーザが応用システムの実態に合わせて維持管理していくためには、ユーザの利用分野に依存する部分はユーザ自身による辞書編集が可能な運用体制が肝要である。このためには以下に示す事柄を考慮すべきであろう。

- ・辞書の提供側との協調体制
- ・辞書編集手順標準マニュアルの整備
- ・辞書編集機能
- ・専門家による技術支援 など

辞書の提供側との協調体制とは、ユーザ自身で管理可能な部分を明確にすると共に、提供側はユーザからのフィードバックおよび自らの改善・拡張を行いユーザに提供していく運用上の協調体制である。図7-2aはこの関係を表わしている。ユーザの辞書変更と提供側の変更によっても辞書全体の整合を保つためには図7-2bで示すように辞書を大きく3つの階層に分け、ユーザは“ユーザ辞書”の層でのみ追加・修正することにより上位層へ影響を与えることなく、それぞれの層の改善に対処できるような運用上の配慮が必要になる。また、これによってユーザは提供側の将来の拡張を受け入れ易くなるメリットがある。なお、実際の運用にあたっては、将来への拡張性を高めるために単に辞書データファイルだけの積み重ねでなく、各層との間に何等かの参照あるいは変換のためのアルゴリズムの存在を検討しておくことも考えられる。

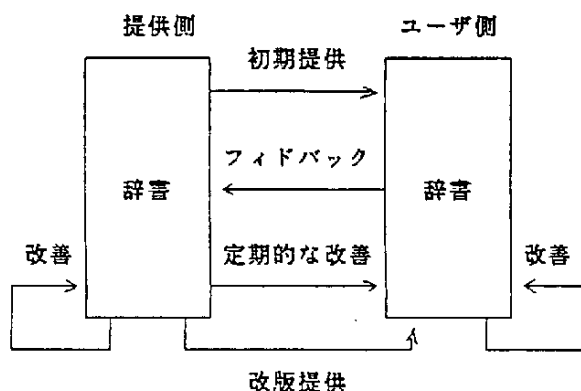


図7-2a

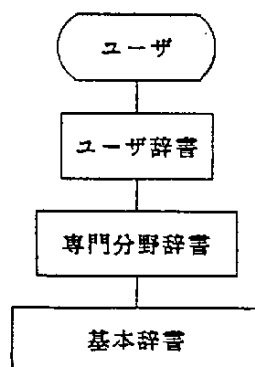


図7-2b

辞書編集手順標準マニュアルの整備については、辞書全体の整合を保持しながら運用管理していくために必須である。マニュアルには、ユーザが辞書のどの部分に、どのような手順と方法により、どのような点に配慮して情報を追加変更していくかの作業基準が記載される。これに併せて作業用のワークシートも用意しておく必要がある。特に辞書編集を一括処理方式で実施する場合には不可決である。

辞書編集機能は、辞書修正をホスト端末機やワークステーションから対話型で実行する機能と、辞書編集ユーティリティにより一括処理方式で実行する機能とに分けられる。対話型機能はいわゆる“辞書エディタ”と呼ばれるもので、適切なガイダンス、マルチウインドウ、参照入力などを用いてユーザが容易に辞書編集出来るための工夫が必要とされる。また、応用システムが対話型で実行される場合には、応用システムから随時に辞書エディタを呼び出せる機能も重要である。辞書編集ユーティリティでは、ワークシートに記載された辞書変更情報を一括入力して、高速に辞書変更するための機能である。いずれの機能においても、辞書全体の整合を保持するためには、辞書変更を誰が、いつ、どのような理由で実施したかの修正の履歴の記録がなされ、かつ、後に検索出来る機能も重要である。更に、未登録語の管理、使用頻度計測、辞書内容の印刷、その他諸統計表の作成機能も辞書を運用・管理していくための支援機能として必要となろう。

専門家による辞書運用支援については、特にエンドユーザに対するシステム運用技術者の支援、システム運用技術者に対する辞書作成技術者などの支援が臨機応変に得られる恒常的な組織作りが必要であろう。また、このような組織における遠隔からの相互のコミュニケーションについても最近急速に進展している通信ネットワーク技術により効率的に実現されるものと思われる。

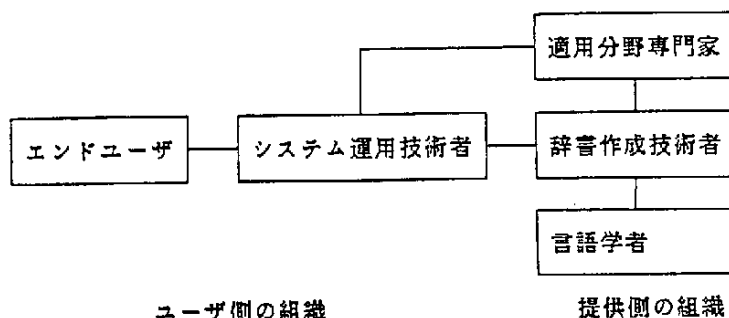


図 7-3 専門家による支援体制

7.3 辞書データ管理システム

実用を目指した適用システムの自然言語解析用辞書の運用には、辞書引きや管理のためのデータ管理システムが重要である。辞書の機能レベルが複雑であったり規模が大きい程、どのデータベース管

理システムを利用して格納・管理するかによってアクセス時間やメンテナンスの効率に大きく影響する。一般にデータ管理機能の充実しているオペレーティングシステムのアクセス方式や汎用のデータベース管理システムに頼って開発の手間を少なくする場合が多いようである。しかし、データ構造の制限やアクセス時間の効率が悪く応用システムそのものの運用にも悪い影響が予想されるので、その選択基準を明確にしておくべきであろう。また今後は、自然言語解析用辞書の運用管理に適した専用のデータベース管理システムの研究開発や更にはニューメディアへの対応の検討も重要と思われる。

7.4 辞書の共用化推進のために

以上、自然言語解析用辞書の運用について、特にユーザの立場から触れてみたが、最後に辞書の共用化に関して更に考察してみる。本来辞書作りは、多大な労力・費用と歳月を要し、かつ言語学者・計算機科学者および応用分野の専門家など多くの人々の協力なくしてはなし得ない仕事である。自然言語解析用辞書作りが生きている自然言語を相手にしている限り、日々、時代と共に変化する言語情報への対応が常に必要とされる意味において、これを次世代への財産として未来永劫に取り組むべき事業と言える。すなわち、一般語を収録する基本辞書については、1研究所・1企業などの枠を越えて分担開発などの協調体制での開発が理想であり、そのための行政の支援が是非とも必要と思われる。願わくば国の事業として永続的に取り組み、例えばそこで開発した辞書の運用に当っては、著作権などの問題もあろうが、基本的にパブリックドメインとして公開し、自由に使える制度とするのが理想的であると言えよう。ただし、公平の原則のために利益を目的とした利用には適切な受益者負担を課するなどの配慮は致し方ないことであろう。同様に分野別の専門用語や固有名詞などを収録する専門辞書についても、公益的に共用可能な専門用語データバンクとして構築されるべく早急に国の機関の設置が期待されるところである。さらに、辞書の共用化を推し進めるためには、言語学者・翻訳家・計算機科学者・辞書作成技術者など辞書作成に関与している専門家の経験・アイデア・提案などの発表および情報交換の場の恒常的な設定も必要となろう。

8. おわりに

コンピュータを用いた自然言語処理、例えば機械翻訳、自動抄録、データベース検索等の各システムにおいて、自然言語の解析を支援するための辞書の重要性についてはすでに衆目の一致するところである。

辞書作成の現状は、各々の目的とする応用システム、さらにはそこで採用している解析レベル、解析手法が異なることにより、辞書の体系、収録項目等にかかなりのバラツキがある。

これは本調査でも各章で述べているとおり、解析のための辞書は自然言語を解析するアルゴリズムやそのための規則と密接な関係を有しているためである。

このため本調査で目的としている相互利用可能な辞書の研究としては、実現困難な点も多くみられる。

しかし、応用システムごとにかから辞書を開発するには、多大な労力、コスト、時間等がかかるうえ、自然言語を対象とする限り辞書開発側で完全な辞書を提供することは困難であり、利用者側にも辞書拡充の負担を強いていることが現状である。

このような状況を踏まえ、直接応用システムで利用出来ないまでも、何らかの加工なり、情報付加すれば利用出来るような、いわばマスター的な相互利用可能な辞書について検討することは非常に有意義なことであり、かつ、将来の利用者ニーズを先取りすることにもなると考える。

辞書の相互利用を意識した場合、配慮すべき事項はこれまで各章で述べられており、工学的、言語学的にみて簡単ではないことが明らかになった訳であるが、今後さらに検討を要する主な課題としては次のような点があげられる。

(1) 辞書作成上の課題

- ① 自然言語解析のアルゴリズム、規則との関係において、解析の各フェーズごとに必要となる辞書の共通項目の明確化
- ② 冊子体型辞書データ（各種一般辞書、シソーラス、用語集、用例集等）の有機的利用方法の検討
- ③ 辞書体系、例えば一般辞書と専門辞書との役割を考慮した全体体系の検討
- ④ 共用可能な辞書項目の範囲の一層の明確化、およびそれに伴う辞書記述のための規則の明確化
など

(2) 辞書運用上の課題

- ① 応用システムで直接利用する辞書（目的辞書）と相互利用可能な辞書（共用辞書）との間の

インタフェース機能の検討

② 共用辞書の維持・管理手法の検討

など

これらの課題解決のためには、一つの応用システムを事例とし特定のアルゴリズムなり規則を基にした辞書を材料として実証的な検討を加える必要がある。

最後に、広範囲に利用する実用的な辞書となると大規模にならざるを得なくなり、かつ、内容的にそれなりの権威が要求されてくることを考え合わせると、辞書の開発、維持、管理体制は民間レベルの枠を超えており、国家レベルの体制作りが急務である。

開発部分については、体制が作られつつあるものの、維持、管理体制は未だ明確ではなく不十分である。

特に、欧州においては分野ごとの専門用語を各国の言語で収録したターミノロジー・データバンクが以前から整備されており、国際取引きにおける活用がなされているが、我が国には未だこのようなものはなく、立ち遅れているのが現状である。

従って、専門用語の標準化も含めた辞書の管理を国家レベルで推進する専門機能の整備が強く望まれるところである。

資料 1 冊子体辞書の歴史と編纂方法

東京都立大学人文学部

忍足 欣四郎 教授

私の専門は中世英文学で、西暦700年～1100年の英語で書かれた文学であり、英和辞典の編纂に当ることになり、以来25年間その仕事に携わった。従って私は辞書編纂の理論より、実務に詳しい。今日は私の専門の英和辞書を中心に、辞書の内容・編纂の歴史・編纂の方法およびいかなる項目が記載されていれば辞書といえるのか、ということについて話す。

まず辞書とは何かということについて述べる。手元にある国語辞典により「辞書」の定義を調べてみた。「三省堂国語辞典」——この辞典では各語は簡潔に定義されている。この辞典のすぐれたところは、用例の多いことである。しかし、「辞典」の定義で「ことばてん」というまだ確立していないことばを使用している。「新潮国語辞典」——辞書と辞典の区別にふれていない。また定義に非論理的なところがある。例えば文字と事柄等レベルの違うものが同列に扱われている。三省堂と異なるのは、定義に用例・語源が入っているところである。「広辞苑」——ここでの定義にもおかしな部分がある。用例は解説されるべきものではなくて、ことばの解説を補助するものである。

以上の例に共通していることから辞書の機能は次のように考えられる。

1. 語を収集する。
2. 収集した語を配列する。
3. 配列した語の説明を加える。

次に、オクスフォード英語辞典(OED)での「辞書」の定義の主要部分を和訳したものを表1に示す。OEDの定義には「辞書」において考慮・検討すべき様々な問題が網羅的に含まれている。最初に収容語彙の範囲を述べている。言葉は歴史を持っているし、地域の変種を持っている。望ましくは各時代、各地域の言葉を網羅的に収集すればよいのだが、冊子体辞書においては不可能なことなので、実用としての辞書はその語彙の範囲を決める。その際そこにある制限を設けなければならない。私は辞書理論の専門家ではないので自信を持って申し上げられないが、語の選択は、おそらく頻度のみでは決められないものと思う。頻度調査の対象のコーパス(corpus)の取り方にもよるが、コーパスをよほど大きくしないと意味のある頻度は出てこないと思う。ごく日常的なことばで使用頻度の非常に高い語とか機能語は問題がない。また、ごく頻度の低い語もまた別の意味で問題がないと思うが、それらの中間の頻度のものについては多少の頻度の多寡は意味を持たないと思う。歴史的原理に基づいた辞書、OEDはそのすぐれた例であるが、すなわち「歴史辞書」というものが有りうる。

表1 「辞書」の定義

三省堂国語辞典	新潮国語辞典	広辞苑
ことばを一定の順序にならべて、発音・意味・用法などを示した書物。辞典。 cf. 「辞典」①辞書。ことばでん。②辞典。…〔多く書名に使う〕	①言葉を集めて一定の順序に並べ発音・意義・用法・用例・語源等をしるした書物。辞典。 ②(略) cf. 「辞典」多くの文字、単語、または種々の事項について、その用字法・発音・意味・用法等を説いた書物。辞書。	①ことばを集め、一定の順序に並べ、その読み方・語義・語源・用例などを解説した書。字書。辞典。辞彙。字引。→事典。 ②, ③(略) cf. 「辞典」辞書①に同じ。
オクスフォード英語辞典 (OED)		
(1) 「ある言語の個々の語(またはその中のある特定の部類の語)を扱って、その綴字法・発音・意味・用法やその同義語・起源・歴史、また少なくともこれらの事実の中の一部を解説した書物。検索に便利ように、語はある一定の順序に、現在ではたいていアルファベット順に、配列されており、また、大型の辞書では、示された知識が文献から取った引用文によって例証されている…」		

また、ある特定の時期(に使用された語)のみを扱った「時代別辞書」というものも有りうる。

地域的に考えると「方言辞書」が考えられる。英語ではブリティッシュ・イングリッシュ、アメリカン・イングリッシュ、旧植民地でのイングリッシュ等がある。

ある文献を読むためという実用的な面から「難語(Hard words)辞典」もある。これは語の実態を明らかにするという学術的見地にたつOEDの目的とはまったく異なり、実用的目的で作成されるものである。

また特定分野で用いられる語を集めた「専門用語辞典」もある。

(以上で一応、辞書の制限ないし、辞書の種類について述べられた。)

再び、OEDにおける辞書の定義にもどり、逐次説明する。

綴字法とは、「どのような綴りを容認されたものとみなすか」ということである。また発音については、「どのような発音を容認されたものとみなすか」と「どのような発音記号を採用するか」が問題となり、発音だけを取り出して扱ったものとして「発音辞書」がある。しかし、なんといっても辞書の中心をなすものは意味の記述であると思われる。現実の辞書の発達を見ると、多くの辞書は先行する辞書の定義(意味の分析)に、多かれ少かれ拠っている訳であるが、理想的には独自に例文を収集し、それらの意味分析を行うことから出発しなければならないし、その分析に基づいて定義を構築するのである。多くの場合一つの語は複数の語義を持つので、その意味を区分し配列することになる。

配列には、歴史的方法によるもの、頻度によるもの、理論的ないし論理的展開によるもの等が考えられる。歴史的配列には、例えばOEDで行っているものがある。頻度による配列は、語義の分析の仕方にもよるので大変難しい。手段としてコンピュータを用いるとしても、頻度を測る文中の語に前もって文脈に沿った語義を与えておかなければならないからである。そうすると、例えばある辞書を

作る場合に、既成の頻度調査は活用しにくいということになる。

以上のような配列の方法があるが、どの配列を採用するかは辞書の使用目的により決定すればよい。常識的に言うと、クーンというアメリカの辞書学者がいるが、この人は中期英語辞典 (Middle English Dictionary) という辞書のエディターであるが、MEDの編集についてある所で次のように語っている。

『MEDはそれ自身を中期英語の厳密に科学的な分析と見なさず、むしろ広く科学の原理に合致する (少なくとも故意にそれを破ろうとはしない) が、多種多様な使用者がなるべく苦勞・混乱しないで求めるものを見い出せるような形で示すように配列した芸術作品と見なしている。』

クーンは続けて次のように言っている。

『論理的配列とは、さまざまな語義、すなわちそれらの語義を示す定義を次のように配列することである。つまり、読者が相互に関連のある、ないし類似したもの (どちらが時代的に後のものであるとか派生したものであるかということと関わりなく) と考えるような語義を分散させずに一括する。語義のグループの中で一番重要なもの (大抵は一番普通に用いられるもの) をグループの最初に置き、また一番重要なグループを語全体の最初に置く。そして、できることなら、どのような語義をも使用者がそれを探そうとしないような位置に置かない、ということである。』

これに関して私が書いた『英和辞典うらおもて』の中で、私の考えを書いているが、このクーンの考えに賛成である。ただし私自身の好みとしては、同一グループ内の配列の原理としては重要度だけでなく、可能な限りで時代的関連や派生関係を考えたい。

定義を構築する場合に重要なことをつけ加えたい。日本語の場合にどうなるか詳しくは考えていないが、英英辞典の場合について、きちんとした辞書はかなり厳密に守っていることがある。とくに動詞など他の語と結びつきやすい語について言えることであるが、『ある動詞に対する定義は特定の文脈で当該の語に代入することができるようなものである』ということである。例えば、他動詞において目的語を採るが、その他動詞に当該定義を代入した場合に、元の目的語がそのまま続くようではなくてはならない。その他動詞を別の他動詞で置き換える場合のみならず、当該他動詞を、例えば (動詞 + 名詞 + 前置詞) のようにいくつかの語で定義したような場合でも、形式的に代入して元の目的語がそのまま続くようではなくてはならない。その点、実用辞書では随分ずさんなものがあるが、OEDにおいては大体、この原則が守られている。もちろん、實際上、定義を代入して英語らしくなるかということは別問題として、論理的・文法的には代入できるように構築されている。

次に、OEDが触れている辞書の側面は「用法」である。用法とは主として統語上の問題であろうかと思う。つまり、語の連結、言い換えると「ある語がどういう語を支配するか」、「どういう成句があるか」、ということであり、このような側面を扱った辞書に「語法辞典」、「慣用辞典」がある。次の側面は同義語・類義語であるが、この側面を扱った辞典が「同義語辞典」、「類義語辞典」である。続いて、OEDでいっている起源・歴史であるが両者はどう異なるのか？ 起源とは、ある語の原

形である。それでは原形はどこまでさかのぼって考えればよいのか？ 英語の場合であると英語の一番古い形から、歴史以前までさかのぼるとゲルマン語・インドヨーロッパ語までを起源という。そういう祖形が歴史の過程での変化をあとづけるものを歴史といい、それをたどって位置づける。そういった側面を扱った辞典が「語源辞典」である。

OEDの「辞典」の定義の最後に、「検索に便利のように、語はある一定の順序に、現在では大抵の言語でアルファベット順に、配列されており」とある。普通の人が使用するものがこれであり、また意味領域でわけた「分類体辞書」がある。ここにいる我々が普通に考え及ぶものとしてはシソーラスがこれに当る。

以上、OEDにおける辞書の定義にそくして、辞書がもつ様々な側面について触れた。

これらの側面は歴史過程において何らかの形で現われてくるので英国における辞書の発達を簡単に述べる。

アングロサクソンは449年にブリテン島に渡ってきた。この年に、ヘンギストとコルサという族長兄弟に率られたジュート族の一団がケントに上陸した。もちろん彼らはラテンもしくは地中海文明に接触していない。彼らはものごとを書きしるす習慣を持っていなかった。ただルーン文字というものを使用していたが、これは金石文用の文字であり、石とか金属、または動物の角、あるいは木片に先の尖ったもので彫りつけたものであり、物事の記録という実用的な用途には用いられず、呪術の意味、あるいは物の製作者の名前をその物に記すことにしか用いられず、事件や文学作品を記述するなどということには用いられなかった。事件や文学的なものは彼らの間では口伝えて伝えられたに過ぎない。以上のような（物事を記述するという）点について、画期的なことだったのは、キリスト教の伝来であった。6世紀末に北と南からキリスト教が渡来する。キリスト教の伝来は、宗教上の出来事にとどまらず、彼らにとって、地中海文明と接触するということであった。そしてアングロサクソンはローマ文字を取入れて記録することを始めた。当時ラテン語はヨーロッパにおける国際的共通語であった。当時は文献というのはラテン語の文献であった。彼らはラテン語の写本を読む場合に、当然のことながら難しい言葉に行き当たることとなり、自分で写本を読むのに役立つように、あるいは、後の人がその写本を読む時の参考になるように、もっと平易なラテン語や、自国語つまり古期英語で、その写本に注解を加えた。一般的に、あるラテン語のテキストについて行間に書き込んだ注解は、難しい語になされている場合と、ほとんどすべての語になされている場合と両方ある。そういう注解を一ヶ所に集めたら便利であろうということで、注解集 (glossary) が作られた。これが辞書の第1歩である。この注解集における語の配列は様々であり、1番原始的な形はそのテキストに出てくる順序に配列するというものであった。少し進んだ形は、単語の最初の1文字（例えばA）の同じ語をひとまとめにするという方法であり、第2文字目以下は問題にしない。もう少し進んだ段階では、先頭2文字による配列が用いられた。第3の配列法としては、主題別に関連のある語を集めたものがある。これを類語集 (vocabulary) といい、古期英語の頃から中期英語の時代に至るまでいくつか

作られている。15世紀になると英国最初の英羅辞典が作られた(1440年頃)。また、羅英辞典が作られた(1460年頃)。ここに本格的な辞書が成立した。その後、文芸復興の新しい気運に乗って、16世紀には古典ラテン語の辞書の編纂が行われ、そしてまた、ヨーロッパのフランス語、イタリア語、スペイン語等を英語で説明した辞書が成立した。

そのようなわけで、英国における辞書は外国語、つまり当時の国際語であるラテン語を対象とした辞書から始まり、近代の英語辞書、つまり国語辞書—英語を英語で説明した辞書—の成立はもっと後になる。

16世紀の中頃になって最初の英英辞典が作られる。しかし、初期の英語辞典は難語(Hard words)の辞書であった。そういう難語辞書の編集は18世紀にまで及ぶわけであるが、一般的・標準的辞書は18世紀に発達することになる。現在までつながる本格的な辞書が作られるのは19世紀のことである。その背景としてドイツを中心としたヨーロッパにおいて歴史言語学の発達があり、歴史的な原理に基づいた辞書として前述したOEDが登場する。

日本の場合、多少英国の場合と事情が似ている。日本はまず中国語から発して、その中国語を日本語で説明する、つまり、漢和辞典を作った。そして、その漢和辞典には部首により分類したもの、あるいは意味表示により分類したものがある。国語辞典が登場するのは、その後であるということで、英国の場合と似ているわけである。

最後に編纂の方法、その具体的手順、手法などについて述べる。辞書において最も重要なのは、意味と用法であると思われるので、これらについて私の編纂した岩波新英和辞典に就いて簡単に述べる。

岩波新英和辞典の「はしがき」にも書いたが、この辞典の性格は、

『英語の学習に強い熱意をいだいている高等学校や大学の学生諸君にも大いに助けになるとともに、英語の知識を必要とする各界の第一線に立つて活躍する実務家や我々英語教育に携わる者にとって、これ1冊あれば少なくとも当座の用が足りるという小型の英和辞書であり、鞆の中に忍ばせて車中で書物や新聞・雑誌を読む時に使用でき、または、教室で英語をするに値する小型の英和辞典』

ということを図示し、簡単に言えば、技術家や教師を対象にした小型の扱い易い英和辞書ということである。使用目的や対象により、見出し語の選定が決定されるわけで、この辞書は、前述の性格のものであるので基本的には新しいものでなくてはならないし、新しい語も含まねばならない。同時に、教室で使うということから現在の文献だけでなく多少古い文献を読む時にも使えるように考えた。そこで重要語の扱いを詳しくしようと考えた。ただし、英語は我々にとって母国語ではないし、語の頻度調査を行ったわけでもない。また、他の所で行った頻度調査を一方的に取り入れるわけにもいれない。そういうわけで我々は、権威ある、しばしば改正されup-to-dateになっている辞書、あるいは新たに刊行された辞書を語彙選定の参考資料として複数選び、どの辞書とどの辞書に収録されているものは取り上げる、という選定方法を採用した。語の配列については、本来ならば多数の例文を収集し、それらにより語義の分析を行い語義を決定し、一定の方法に従って配列するわけであるが、こ

れも外国語であるのであちらの辞書に頼らざるをえない。そして、当初はThe American Heritage Dictionary of the English Language (AHD)を第一の基礎資料とした。その理由の第一は、AHDがコンピュータを利用して編纂されていること、その第二は、米国で出たごく新しい辞書であるということである。しかし、AHDを資料として使用しているうちに、この辞書がまったく信頼できないものであることがわかった。そこで、校閲者に、常にOEDを参照して修正するように指示した。今日では辞書において意義素的な分析が重要視されているが、OEDは、今から100年以上も前に、編集が開始された時点で、すでにそのようなことを行っていた。

表2に、例としてOEDにおける副詞“Round”についての意義分析を示した。定義を日本語に直し、簡単な例文を添えて、意義の要素を示し、図式化したのが表2にあるものである。

“Round”の基本的意義は「円」であるが、OEDはそれを大きくI、IIと2つに分けている。OEDはどういう基準でIとIIに分けたかは書いていないが、それぞれにおける定義を逐次読んでいくと、Iは「公転」すなわち回転するものが回転軸をその内に持たない場合、IIは「自転」すなわち回転軸をその内に持つ場合、であることがわかる。表2において「意義の要素」は、運動の観念が含まれるものには「+運動」、含まれないものには「-運動」で示し、また、360°閉じている円の観念のものには「+完全」、完全に閉じていないものには「-完全」で示した。

もちろん完全なものでないにしても、この時代にOEDはすでに語義の分析を行っていた、ということがわかる。

先程も述べたように、OEDは歴史主義（語義の生起した年代順に配列していく）であるが、意義の分析を行っているように、いわゆる論理主義と歴史主義とは必ずしも矛盾するものではない。つまり、一つの語義を認定するのは難しいことではあるが、OEDにおいては、語義番号1, 2, …の中では、一応年代順になっている。ただし下位の語義番号の間では必ずしも、年代順にはなっていない。例えば、1-b (1613年)は2-b (1836年)よりも初出が新しい。このようにOEDにおける語義の配列は歴史主義と論理主義の組み合わせになっていると考えてもよい。以上のOEDにおける“Round”の定義を分類してみたのが、図に示したものである。左側の分類は、公転/自転によりまず分類してみたものであり、右側は、まず運動の観念があるかないかにより分類してみたものである。

以上見てきたように、私は、現在刊行されている辞典の中でOEDほど秀れているものはないと考えている。もちろんOEDにも様々な欠陥はあるが全体として見るとOEDほど秀れているものはない。このような理由で英和辞典作りに際しても、OEDを大いに参照させてもらっている。

表2 OEDにおけるRound adv. の意義分析 (基本的意義は「円」)

語義番号		定義 (初出年代)	例 文	意義の要素		図 式
I 公 転	1	ぐるりと回って、一周して、循環して；(時が) めぐって a 1290	go ~ and ~ in a circle ぐるぐると輪を描いて回る / Summer will soon come ~ again. 間もなく夏が再びめぐって来る	+運動	+完全 (360°)	
		b (一巡して) 皆にわたるように 1613	There is just enough cake to go ~ ちうど皆にわたるだけの菓子がある			
		e 初めから終りまで；ずっと，…中 1753	all the year ~ 一年中			
		f 巡回 (巡遊，巡訪) して 1821	show sb. ~ 人をあちこちに案内する			
		g 動き回って 1890				
	2	ぐるりと囲んで；まわりに a 1290	a room hung ~ with portraits 壁にずらりと肖像画をかけた部屋	-運動	+完全 (円周)	
		b 輪 (円形) をなして c 1386				
		[周囲が]	a log 3 feet ~ 周囲3フィートの丸太			
	3	(中心から) あらゆる方向に，到る所に，あちこちに c 1440		-運動	-完全 (円周上の点)	
		b (中心から) 四方に向かつて測って 1650	all the neighbors for a mile ~ 1マイル四方の全住民			
		c 近くに，辺りに 1785				
	4	回り道して 1668	go a long way ~ 随分回り道する	+運動	-完全 (<360°)	
		b 自分のいる場所へ [から]，出向いて 1698	ask sb. ~ 人にこちらへ来るように言う / Order the car ~ 車を (玄関) に回させてくれ			
II 自 転	6	ぐるぐる回って，回転して c 1500	Wheels go [turn] ~ 車輪は回転する		+完全 (360°)	
	8	反対側に，反対方向に a 1765	He turned ~ 向き直った / Don't look ~ . 振り返るな		-完全 (180°)	
		b 反対の意見 (態度) に 1825	talk sb. ~ 人を説得する			

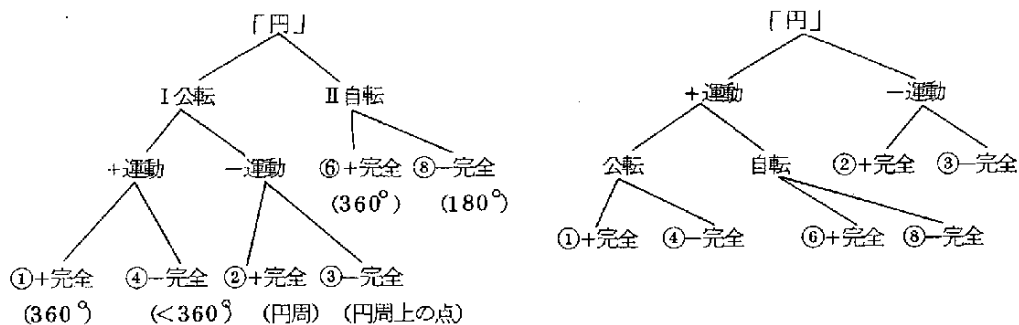


図. “Round” の定義の分類

〔質疑応答〕

Q 1. 今までの辞書（例えば OED）の歴史的変遷をお話いただいたが、これからの辞書はどうあるべきか、今までの辞書の不足している点はある所か、それはとりもなおさず辞書の未来像となると思いますが、いかがお考えでしょうか？

A 1. 冊子体辞書を考える限りは、意義素的な分析をより徹底しておこなうこと。それから、これは英語辞書の場合であるが、学習辞典にあるような文型の表示（例．SVO など）や countable / uncountable の区別等についてより厳密・体系的に理論的な究明を加える必要がある。後者に関する動詞を中心とした分析は、Longman の LDOCE では非常に詳しく行われている。OED においては、必ずしも体系的ではないがかなり細かく分析されている。

Q 2. 「意味」と「意義」の差異は？

A 2. 表 1 にある限りにおいては差異はない。詳しくは、使用される文脈まで考えねばならない。

Q 3. 「用法」と「用例」の差異は？

A 3. 「用法」…主として語と語の結びつきなど、統語上の問題点を示す。

「用例」…項目をなす言葉がどのような文で使用されているかを示す。

Q 4. 語釈はどこまで詳しくすればよいのか？

A 4. 語釈において使われる言葉を見出し語にすべきか、については、理想的には現出する語はすべて載せるべきであるが、現実には難しい。

Q 5. 「読み方」と「発音」の差異は？

A 5. 「読み方」とは漢字の読み方が念頭にあるのではないか。

Q 6. 頻度調査の問題点はなにか？

A 6. ①中間頻度を持ったものに対しては、頻度調査を絶対的なものと、とらえることはできない。

②文学作品の中に出てくる語などで、頻度の低い語でも、採用してもよい場合があるのではないか？

Q 7. 独自の語義分析というものは、どのくらい行われているのか。

A 7. 英和辞書ではほとんどすべてが先行辞書の引用である。英英辞書でも、先行辞書に配列を変え
るとかして、少し手を加えた程度の場合が多い。

Q 8. 新語を考えた場合、見出し語の選定は先行辞書に頼ってばかりいられないと思うが現実にはど
のように行われているのか？

A 8. エディターにもよるが良心的辞書の場合は選定作業を行っている。また新語辞典というものが
あるが、私の辞書もそれを参考にしていて。新語または語の新しい用法について（何しろ外国語
のことであるので）その意味を定義するのは自信がもてないし、独自に定義するのは難しい。ま
た、その新しい語の将来の使用頻度や用法の見極めが難しい。

Q 9. 現在の辞書作りに、言語哲学あるいは言語学が反映されているのか？

A 9. いろいろな辞書を見ても反映されていない。

①OEDでの厳密な語義区分の方法について、英国で編集当事者に聞いても、日本の英語の語
選定の専門家に聞いても、「方法」は知らないと言っている。OED全体としてみると語義
区分は一貫している。従って内部資料はOED関係者に渡っていると推察する。

②日本の国語辞書でも欠陥が多い。辞書学の理論と辞書作りの現場は結びついていないと聞い
ている。

Q 10. 岩波新英和辞典はOEDを参考にしているようだが、図にある“Round”の体系図のような
構造化したものを作成されたのでしょうか？

この体系図は我々のプロジェクトに非常に参考になります。

A 10. この体系図は私の頭の中で考えたものである。OEDではマレー氏のような優秀な人が編集の
指揮をとっていた（ので、このような体系的な語義区分が達成できたのだと思う）。「岩波新英
和」ではOEDそのままではなく、私自身の判断で考えた部分がある。OED本体は古い時代に編

集されたものであるし、また長考の末、米語も採り入れることにした。

Q 11. 一般論として辞書改訂には長時間かけるべきか？

A 11. 理想的にはどんどん改訂していった方がよい。言葉によって、変りやすいものと変りにくいものがある。全体としては、辞典の性格によるのではないか。例えば学習辞書は新聞・雑誌を読むために使う用途ではないので余り改訂する必要はないと思う。

資料 2 ICOT における辞書の研究について

新世代コンピュータ技術開発機構 (ICOT)

研究所第二研究室 横井 俊夫 室長

1. 動機

何故、第 5 世代プロジェクトで電子化辞書を作成しようとするのか？

それは、第 5 世代プロジェクトの中で、知的インタフェースというテーマに一応くられているが、自然言語処理は知的処理の根幹にかかわるテーマである。しかも、自然言語処理というテーマにおいても、単に個々のシステムの開発のみではなく、本格的使用に耐えうる第 5 世代技術の 1 セットを作ろうとしている。したがって、そのための基礎ともなるべき辞書も、中途半ばなものではなく、本格的辞書作りを志向することになった。

2. 経緯

第 5 世代で電子化辞書を作成することになったが、数万語も初めから収集するのは困難であることから、冊子体辞書の出版社が所有している辞書情報を元に、電子化を行うことになり、数社の出版社に当たってみた。固辞すく出版社もあったが、情報の電子化の流れの中で大きな変化を受けるのは出版社であろうと言われている折でもあり、積極的姿勢の出版社もあり、コンピュータのメインフレームと共同で電子化辞書を開発する体制が整ったのが昭和 59 年 7～8 月頃であった。

3. 内容・体制

3 年間で、以下に示す体制により各 5 万語の電子化辞書を計 4 本作成することになった。1 本当り 3 年間の開発費は約 2 億 5 千万円であるという計画となった。

辞 書	メインフレーム	出版社
1. 日 本 語	日 立	(未 定)
2. 英 和	東 芝	学 研
3. 和 英	富 士 通	旺文社
4. 英 語	富 士 通	旺文社

また、ICOT の電子化辞書を検討し推進するため、茨城大学教授であり、言語学者である石綿敏男先生を委員長とする委員会を設置した。

4. 移行

第5世代プロジェクトは予算が大きいと一般に考えられているが、テーマ数も多く、辞書作りに割当てられる予算は十分ではなかった。

そこで別の形態、予算枠での新しい研究・開発事業として計画が進められている。この計画が決まれば、I C O Tにおける

辞書作成事業は昭和60年9月をもって終了することになる。

ちなみに、I C O Tでの辞書作成は、(最重要語1千～2千語+シソーラス)の辞書と、逐次型推論マシンP S I上での辞書作成支援ソフトウェアの作成である。新しい事業となれば、辞書の規模も、もう一段拡大し、概念体系、概念記述等の高度な研究も含まれることになる。

5. 私見

プログラム言語は、万国共通の言語でよいが、情報処理技術は出来る限り、それぞれの国の国語で表現できるようにしたい。例えば、ドキュメントコメント、変数名、等は日本語で表現すべきであろう。

逆に言うと、日本はガソリンが無くて困ったと同様に、“ことば”で苦労しているが故に、“ことば”に対する技術こそ日本の有力な武器になりうると考える。

(質疑応答)

Q 1. 電子化辞書の現在の問題点は？

A 1. 語彙数から、中の仕組から、すべて見直す気運にある。一般用語の語彙数も増したいと考えている。

Q 2. 最重要語1千～2千語の選択はいかにするのか？

A 2. 各メーカーにまかせた。例えば、冊子体辞書で星印* *がついている語、等が考えられる。

Q 3. 「語」の例を示してくれ？

A 3. 昭和59年度I C O T報告書に掲載されている。また、今年10月末に磁気テープ化された辞書が出来上がる。

Q 4. 一般に、辞書を引くのは、その語がわからないからである。その意味からも、国としては、専門用語辞書の開発を重視すべきであろうと考えるが、いわゆる基本辞書と専門用語辞書の切り分けはいかに行っているのか？

A 4. 基本用語は、現行の(冊子体)辞書を生かすことにしている。

機械翻訳等の自然言語処理技術からの要請が研究開発の動機であるから、むしろ基本語の方が

より重要である。

Q 5. シソーラスは開発するのか？

A 5. 昭和60年9月までに開発したシソーラスは、あくまで参考的なものであり、今後3年間でより本格的なシソーラスを作る予定である。

Q 6. 専門用語辞書作りに必要なのは、どういう分野の人か？

A 6. 言語学者、情報処理技術者、目的とする分野の専門家である。

Q 7. 新しい事業で、作る予定の辞書の提供形態は？

A 7. 磁気テープによる提供をまず考えている。

Q 8. 辞書の規模を5万語としたのは何故か？

A 8. 携帯用（冊子体）辞書の収録語数をメドとした。

資料 3 機械翻訳システムの評価と利用法

電力中央研究所

寺野 隆雄 研究員

1. 動機

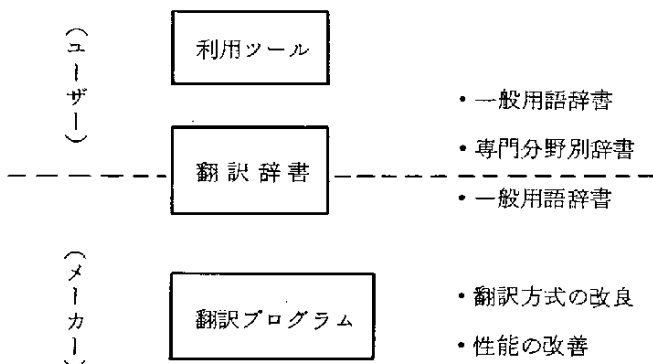
財電力中央研究所では、大量の技術文書の翻訳作業を省力化・効率化する必要があることから、英日機械翻訳システム（富士通 A T L A S - I）の評価実験を行い、現在の技術レベルで適切な利用方式を探った。

2. 機械翻訳システムの特徴

現行の機械翻訳システムには、種々のあいまいさを解決するための自然言語処理技術上の制約や、システムの使価や C P U 時間等のコスト上の制約などがある。ユーザ自らがその能力と限界について十分な認識を持ってシステムを使いこなしていくことが重要である。

今後数年間は、機械翻訳システムを翻訳の支援に用い人間が機械翻訳文を修正して読みやすくするという使い方が中心となろう。

その場合、とくに専門分野別辞書の充実が翻訳性能に深くかかわっている。



機械翻訳システムの利用体制

3. A T L A S - I による機械翻訳の実験とその評価

以下の条件で評価実験を行った。

- ・対象システム： 英日自動翻訳システム A T L A S - I (基本辞書 5 万語, 文法 7 千)
- ・例題英文： E P R I Journal (米国電力研究所) の記事アブストラクト 12 編
- ・評価尺度：
 - 英文 (入力) の複雑度 → Fog - Count に基づく尺度
 - 日本語 (出力) の読み易さ → 理解容易性 (5 段階)
 - 日本語 (出力) の複雑度 → 忠実度 (7 段階)
 - 翻訳コスト
- ・翻訳手順： 次の①～④の 4 つの手順による翻訳結果を評価した。
 - ①入力テキストの一括翻訳
 - ②入力テキストと翻訳用辞書の対話修正
 - ③編集済み辞書による一括翻訳
 - ④人手による翻訳 (比較用)
- ・判明事項：
 - ①機械翻訳システムは一応利用可能なレベルに達している。そして原文の対象分野にあわせて翻訳用辞書を整備することで比較的容易に訳文の品質を向上させることが可能である。
 - ②人手で原文の入力を行う限り, その作業時には必ずミスが発生する。また, 原文自体につづり違いや文法的不備など, さまざまな誤りが存在することも多い。したがって機械翻訳システムを実用化するには, このための対策を十分考慮しておくことが重要である。
 - ③ A T L A S - I の対話処理機能は改善の余地が多い。それは大型計算機の T S S 環境の操作性の悪さと, 翻訳実行速度の遅さの 2 つの問題に起因する。対話処理の操作性がワード・プロセッサのレベルまで改善されれば, 現状のシステムでも一般利用者が十分活用できるものと考えられる。
 - ④ A T L A S - I の辞書編集機能は, 既存の機械翻訳システムの中では非常に優れている。しかし, それでも, 辞書の追加や修正の作業はかなり難しい。A T L A S - I で採用された機械翻訳の基本方式を理解することはもちろんのこと, それに加え, 対象分野の専門知識; 計算機利用全般に関する知識; 英語ならびに日本語の素養が必要となるためである。

4. 機械翻訳システムを導入する場合の留意点

上記留意点として, (i) 利用者が実際に翻訳したい文章によるシステムの評価を行うこと, (ii) システムの利用者層・利用目的の明確化, (iii) 基本辞書・専門用語辞書の整備によるシステムのチューニング, (iv) 利用体制の確立 (翻訳利用者とシステム管理者の分担) 一が指摘された。

また, 再三にわたりその重要性が指摘された専門用語辞書の整備手順についても説明があった。

①原文の分析 → ②専門用語・抽出分類 → ③専門用語・訳語付け → ④専門用語・翻訳辞書化 → ⑤専門

〔質疑応答〕

Q 1. 商用機械翻訳システムでユーザがさわれる辞書の編集にも専門家が必要とあるが、この場合の専門家の範囲は？

A 1. 当該専門分野の用語がわかり、ある程度翻訳システムの原理を理解している人であり、言語学者とは異なる。

Q 2. 辞書のさわり方は？

A 2. メーカー提供の辞書は一般ユーザは不用意にいじらない方がよい。ユーザ辞書についても辞書によりさわれる者の資格を設ける必要がある。

Q 3. 同じ文法カテゴリーでも意味が違う場合はどうするか？

A 3. このシステムでは、1つのカテゴリーには原則的に1つの訳語しかつけられない。専門用語辞書の切替え（具体的には引いてくる優先順位を変える）で対応するしかない。現在のところ翻訳コストを見ると、人手による方が一次入力文よりコスト的に良い。しかしこれは、初期の日本語ワープロに見られた状況によく似ており、現時点でシステムを導入してもいちがいに不利とは言えない。それは次の理由による。第1にシステムによる方が翻訳スピードが速いこと。第2に、2～3年すればシステムのハードウェア価格が低下すると考えられること。

Q 4. 本文と要約文では複雑度が異なると考えられるが比較したか？

A 4. 比較は多数行い、複雑度に大きい相違は見られなかった。

Q 5. 専門辞書の分野分けにおいて、テーマとしての分野とカテゴリーとしての分野をいかに区別しているか？

A 5. 区別することは重要である。しかし、その分類方法は現在のところ見えていない。

Q 6. Muプロジェクトでは、語の多義性を回避するため、語に分野コードをつけて、初めに利用者の分野を向うというシステムの作り方を志向しつつある。そうすると、“分野”とは何かが問題となり、文部省の専門用語分類作業と並行して進めねばならない。

ATLASでは、1語1訳を改良する予定は聞いているか？

A 6. ATLAS-Iは、辞書引き1回なので原理的に不可能であろう。

Q 7. 機械翻訳システム改善で、どういう所（高級または泥くさい）に力を注ぐべきか？

A 7. 泥くさい所に力を注いでほしい。（例：“as” = “時” または“比喩”）

原理は簡単でも手間をかけた方が良い翻訳が得られる。

自然言語解析用辞書

アンケート調査票

財団法人 日本情報処理開発協会

目 次

1. プロジェクト及び辞書システムの概要
2. 辞書の概略及びその内容について
3. 辞書の作成について
4. 辞書の運用について
5. 辞書の他のプロジェクト及び他のシステムとの
共用使用について
6. その他

フェイス シート

アンケート調査にご協力いただき、ありがとうございます。調査に御回答くださる前に
以下の事項の記入をお願いいたします。

1. 記入年月日 _____
2. 貴社及び機関名 _____
3. 所在地 _____
4. 所属 _____
5. 御芳名 _____
6. 御連絡先（電話） _____ 寓
7. その他、ご意見など

I. 参考各

1. プロジェクトについて

プロジェクトの正式名称	
プロジェクト参加機関名 (複数の場合)	
プロジェクトの性格・目的 (例) ○○の助成金により、○○の確立を目指す、○○との共同研究	

2. システムについて

* 当該辞書を使用しているシステム (自動翻訳システム等) についてお答え下さい。

(1) システムの正式名称 _____

(2) システムの種類

a) _____ (自動翻訳、データベース作成・検索等)

b) システムの目的 (実験システム、商用システム、その他) _____

c) システムの特徴について簡単に記入して下さい。

(例) 毎時○万語の英日翻訳を○○%の変換精度で行う。

d) 使用計算機システム

機種 _____

OS _____

開発プログラミング言語 _____

e) システムの大きさ

データの容量 _____

プログラム容量 _____

その他 _____

f) システムの機能概要（機能図）を記入して下さい。

II. 機械可読辞書について

1. 機械可読辞書の正式名称 _____

2. 機械可読辞書の作成日（作成期間） _____

3. 機械可読辞書の種類

（既存の冊子体辞書の機械化読化、機械翻訳システムの解析用辞書 等） _____

4. 機械可読辞書の種別（日本語、和－英、英－和、英語 等） _____

5. 辞書の内容について

a) 機械可読辞書の現時点での対象分野ごとの収録単語件数及び、今後の予定件数について記入して下さい。

	分 野	収録単語数	予 定 件 数
一 般			
専 門			

b) 収録単語の今後の予定について記入して下さい。

6. データ構造について

* a) ～d) までについてお答え下さい。記入は、p 5 ～p 8 の記入欄をお願い致します。

a) データ項目

当該機械可読辞書のデータ項目、及び各データ項目の機能・目的等について別紙1・別紙2を参考にして記入して下さい。

b) 各データ項目の内容構成

各データ項目の内容構成について、その詳細を別紙3・別紙4・別紙5を参考にして記入して下さい。例：①見出し語に対する収録語の種類

②品詞分類

③意味マーカ、及びその体系

④格フレーム、及びその体系 など

c) 各データ項目のデータフォーマット

各データ項目のデータフォーマットを記入して下さい。

d) 辞書のファイルフォーマット

当該機械可読辞書のファイルフォーマット、及び各データ項目間の論理構造について、別紙6を参考にして記入して下さい。

a) データ項目

b) 各データ項目の内容構成

c) 各データ項目のデータフォーマット

d) 辞書のファイルフォーマット

Ⅲ. 辞書の作成について

1. 機械可読辞書の作成及びメンテナンス体制

a) 機械可読辞書の作成及びメンテナンスについて、どんな体制を敷いていますか？

〔例〕 出版社と共同、入力についてはA社と共同など

b) 当該機械可読辞書の作成体制について、以下の項目についてお答え下さい。

人員構成 (人数・資質)	〔例〕 言語学専門家 ○○名 情報処理技術者 ○○名 インプットオペレータ ○○名 ○○○○ ○○名 ○○○○ 外注	
工数 (人月)		
経費		

c) 当該辞書のメンテナンス体制について、以下の項目にお答え下さい。

メンテナンス内容	
頻度	
メンテナンス手順	
人員構成（人数・資質）	
経費	

2. 辞書の作成手順

当該機械可読辞書について、どのように作成しておりますでしょうか？
その作成作業の手順の概略を記入して下さい。

3. 収録単語の収集方法

当該機械可読辞書に収集されている単語は、どのようにして収集していますか？その収集方法についてお答え下さい。

4. 収集単語の分類作業

a) 収録する単語の収集から分類、各種マーカの付与に至る作業手順の概略を記入して下さい。

b) 収録単語の選定基準について記入して下さい。

5. 機械可読辞書管理システム

当該機械可読辞書の作成・メンテナンス作業（入力、追加、修正、訂正など）を行うシステムについて、その機能概要を別紙7を参考にして記入して下さい。

6. 入力方法

当該機械可読辞書の作成にあたり、入力作業に於けるデータエントリー・ルールを記入して下さい
また、参考となる辞書作成マニュアル、作業フォーマット（別紙8参照）等サンプルがありましたら、添付して下さい。

7. 機械可読辞書の作成・メンテナンスに於ける方法論、及び今後の作成・メンテナンス体制についての予定、問題点、意見等をお書き下さい。

IV. 辞書（システム）の運用について

1. 主なユーザについて

- a) 社内、商用、その他 _____
- b) ユーザ数 _____
- b) 主なユーザの使用分野 _____

〔例〕電力関係、原子力関係、通信関係

2. システムの利用に於けるエンドユーザ側の機械可読辞書の操作可能部分と操作方法

a) 辞書の登録・修正・改良に関して、どのような体制になっていますか。

(エンドユーザに任せている・エンドユーザと協力体制を敷いている・その他)
(エンドユーザが操作可能の場合、エンドユーザ用に提供しているマニュアルがありましたら添付して下さい。)

b) エンドユーザによる、辞書の修正（書き換え）・追加可能部分及びその方法についてお答え下さい。

修正・追加可能部分	そ の 方 法

c) システムの辞書以外の部分で、エンドユーザが操作可能な範囲についてお答え下さい。

d) エンドユーザによる、辞書の実際の操作内容について、その状況を具体的にお答え下さい。

e) エンドユーザによって辞書が追加・修正・訂正された場合、その権利（著作権 等）はどのように扱われますか。

V. 当該辞書その他プロジェクト、及び他システムとの 共用使用について

1. 当該機械可読辞書を他のプロジェクトあるいは他のシステムと共用使用する場合、どのような問題が生じると思われますか。

2. 共用使用を前提とした場合、機械可読辞書の作成及びメンテナンスを行う上で、考慮すべきだと思われる点があればお書き下さい。

3. 提供可能性、及び提供方法について
 - a) 当該機械可読辞書が、有償あるいは無償で外部に提供される可能性はありますか。

 - b) 提供される場合の条件がありましたら、お答え下さい。

 - c) どのような媒体で提供されますか。

 - d) 有償の場合、提供価格はどの程度とお考えですか？

VI. 関連資料、参考文献等がありましたら御紹介
下さい。

—— 禁 無 断 転 載 ——

昭和 61 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) 8 2 1 1 (代表)

印刷所 三協印刷株式会社

東京都渋谷区渋谷 3 丁目 11 番 11 号

TEL (407) 7 3 1 6

