自然言語処理 38－7
（1983. 7. 15）

語の収集と体言を中心とする辞書について

中井 浩　佐藤 雅之
（日本科学技術情報センター）

1．はじめに

科学技術文献を翻訳の対象とする場合には、専門用語等の名詞を主体とする体言辞書の果たす役割は大きい。体言辞書を作成するときに、以下の問題点がある。

1)量的問題…名詞を中心に、範囲、収集、選択、即ちカバー率が問題になる。
2)語単位の問題…特に、専門用語の語単位の同定が問題になる。
3)対訳語の問題…語単位の問題に関連し、特に長単位語の英訳語の表現が問題になる。

本編では、以上の問題に言及するとともに、主として体言辞書として持つべき情報、辞書作成作業上の問題点について述べる。

2．語の集収方針

2.1 量的問題

範囲を科学技術分野に限定して、専門用語集の数を調べたところ、数百種にのぼることがわかった。[1] また、日本科学技術情報センター（JICST）でデータベース作成の参考資料として使用している用語集を調査した結果によると、日英専門用語集は約150種にもなることが確認された。

JICSTで用いている科学技術分類をもとに各分野の代表的な用語集32種を選んで調査した結果、延べ語数は約50万語に達した。また、JICSTデータベースの文献タイトルから切り出した語について調査した結果、約87万件の文献タイトルに出現する異なり語数は約80万語であった。

科学技術の専門用語は、用語集及び用語の数とも極めて多い。翻訳辞書には大量の実用的な専門用語を収録することが必要であるが、数10万語から100万語の辞書を短期間に作成することは困難であり、必ずしも実用的でない用語も多く含むことになる。このため、科学技術文献に出現する専門用語を調査し、効率的な辞書作成手順を検討した。

(a) 文献タイトル中の専門用語

JICSTデータベース中の電気工学分野の約5万文献のタイトルを調査した結果、異なり語数は約67,000語であった。出現頻度別の語数は下表の通りである。

第1表 出現頻度別語数

出現頻度	切り出し語数	割 合
1	51,487語	76.6%
2	7,059語	10.5%
3	2,527語	3.8%
4	1,259語	1.9%
5	841語	1.3%
6	586語	0.9%
7	424語	0.6%
8	318語	0.5%
9	275語	0.4%
10	202語	0.3%
11以上	2,143語	3.2%

本研究は国の科学技術振興調整費による『日英科学技術文献の速報システムに関する研究』の一部として行ったものである。辞書に関連する部分の研究は、田中康仁氏（姫路短期大学）、石川徹也氏（図書館情報大学）、萬濃立夫氏（日本コンベンションサービス）をはじめとする自然言語処理の専門家により組織された辞書ワーキンググループのメンバーの協力を得て進められた。

タイトルから切り出された語を見ると、頻度1の用語のほとんどが数量詞か長い複合語であり、語基的な用語は高頻度を示していることがわかる。

(b) シソーラス用語

シソーラス用語は，情報検索に必要な標準化された用語であり，検索効率を考慮した用語であるため，通常の科学技術文献中にもよく出現すると考えられるものである。シソーラス用語を任意抽出して文献タイトル中への出現度合を調査したところ，73%が実際に使用されており，高頻度で出現することが確認された。

以上の調査を行った結果，短期間に効果的な辞書を作成するための情報源として、以下のものを対象とした。

①JICST科学技術シソーラス用語 41000語
②JICSTデータベースのタイトルから切り出した語 40000語
③入手可能な既存磁気テープ辞書 161500語

また，名詞以外の体言（副詞，連体詞，接続詞）については，新明解国語辞典より抽出することとした。

2.2 語単位の問題

基本的には前述の既存辞書の見出し及びJICSTシソーラス用語をベースとするが，実際の文章中には短縮語，長い複合専門用語が存在する。特に，抄録文においては，限られた文字数でより多くの情報を持たせようとする為，この傾向は顕著である。これに対処するためには，それらの語を形態素解析レベルで未知語として出現させ，それに対し，辞書側で素早い対応ができるように辞書管理システムを作っておく必要がある。

2.3 対訳語の問題

(a) 上述の未知語に対する訳語付与をいかに行うかが問題になる。また，既存辞書の中には，実世界の中で，もはや変位していて実用語になっていない語が存在する。この修正が問題になる。

(b) 名詞以外の副詞，連体詞，接続詞に対する訳語付けは，慣用表現，イディオム表現，係受けの中で生成する必要があり，問題である。

(c) 複合語における多様な訳語の問題もある。

（例1）
air conditioning ⟷ air-conditining

（例2）
hot air heating
hot blast heating
warm air heating
warm blast heating
} いずれも『温風暖房』

3. 体言辞書情報

機械翻訳用の辞書を構成するにあたり、体言（名詞，副詞，連体詞，接続詞）に与えるべき辞書情報について，実データをもとに分析を行い，表形式で記述することとした。表形式のものは，名詞辞書用と副詞・連体詞・接続詞用の2通りがある。第2表に名詞辞書作業用フォーマットを示す。このフォーマット中の項目で見出し情報と形態素情報は動詞の辞書フォーマットと共通である。構文・意味情報として，分野コード，構文品詞，品詞細分類，変換見出し，意味マーカー，シソーラスコードなどを持つ。さらに共起情報欄により，格構造や意味マーカーのみでは正確に規定できない多くの慣用的表現に関する情報を与えることができるようにした。

体言（副詞・連体詞・接続詞）辞書作業フォーマットは，形態素情報までは名詞とほぼ同じで，構文情報として品詞細分類コードを持つ。副詞に関し

第2表　名詞辞書フォーマット

見出し情報	見出し語			構文・意味情報	分野コード					
		語尾字数			構文品詞	名　詞				
		漢字部			品詞細分類	固有，普通，動作（サ），動作（他），副詞，格助詞，接続助詞，補文標識				
		語基			変換見出し					
		読み			意味コード					
					シソーラスコード					
	異形語				備考					
	派生語			共起情報	形容詞					
	関連語				形容動詞					
形態素情報	形態品詞	名 副名 動 形 形動 副 連体 接 助 助動 接頭 接尾			名詞					
	動詞活用型	五 上一 下一 サ変 ザ変 カ変			その他					
	活用行	ア カ ガ サ ザ タ ダ ナ バ マ ラ ワ			意味コード					
	形動活用型	ダナ ダノナ		備考						
	助動活用型	形 形動 動 特								
	助詞細分類	格 接 副 並 終 準								
	接尾活用型	体 動 形 形動								
	前接情報									

では，さらに意味情報を付与する。

副詞のプロパティーとして以下のものを設定した。

① Modality（仮題，疑問，打消，願望，比況，譲歩）
② Aspect（完了，反復・習慣，進行）
③ Tense（過去，現在，未来）
④ Gradability（尺度，極）

3.1　品詞細分類

辞書における品詞の細分類を第3表に示す。この表は，辞書情報として持たせる品詞を1つ通りに大分類し，それぞれを細分類して記号と用例を与えたものである。

細分類品詞のうち「体述的形容動詞」だけは，辞書情報として持つのではなく，形態素解析の結果でつける品詞である。句接続詞は文と文，句と句のどちらも接続できるのに対して，文接続詞は文と文のみの接続にしか用いることができない点がちがう。

3.2　意味マーカー

名詞の意味マーカー（意味素性）として，第4表に示すようなファセット構造を持たせたものを設定した。意味マーカーは2字の英字で表現され，1つの名詞に対して，最大5個まで付けてよいことにしてある。

意味マーカーを付与するときは，一般的に以下の事が言える。

① 接尾語がついた複合語は，その接尾語によって意味が決まり易い。

例.
安全性 ⟶ ＡＰ（性質）
標準化 ⟶ ＭＳ（基準・標準）

② 派生語は，最後の主語基によって意味を判断する。

例.
アンケート調査→調査→ＤＡ(行為)

名詞の意味マーカー付与作業の流れ図を第1図に示す。

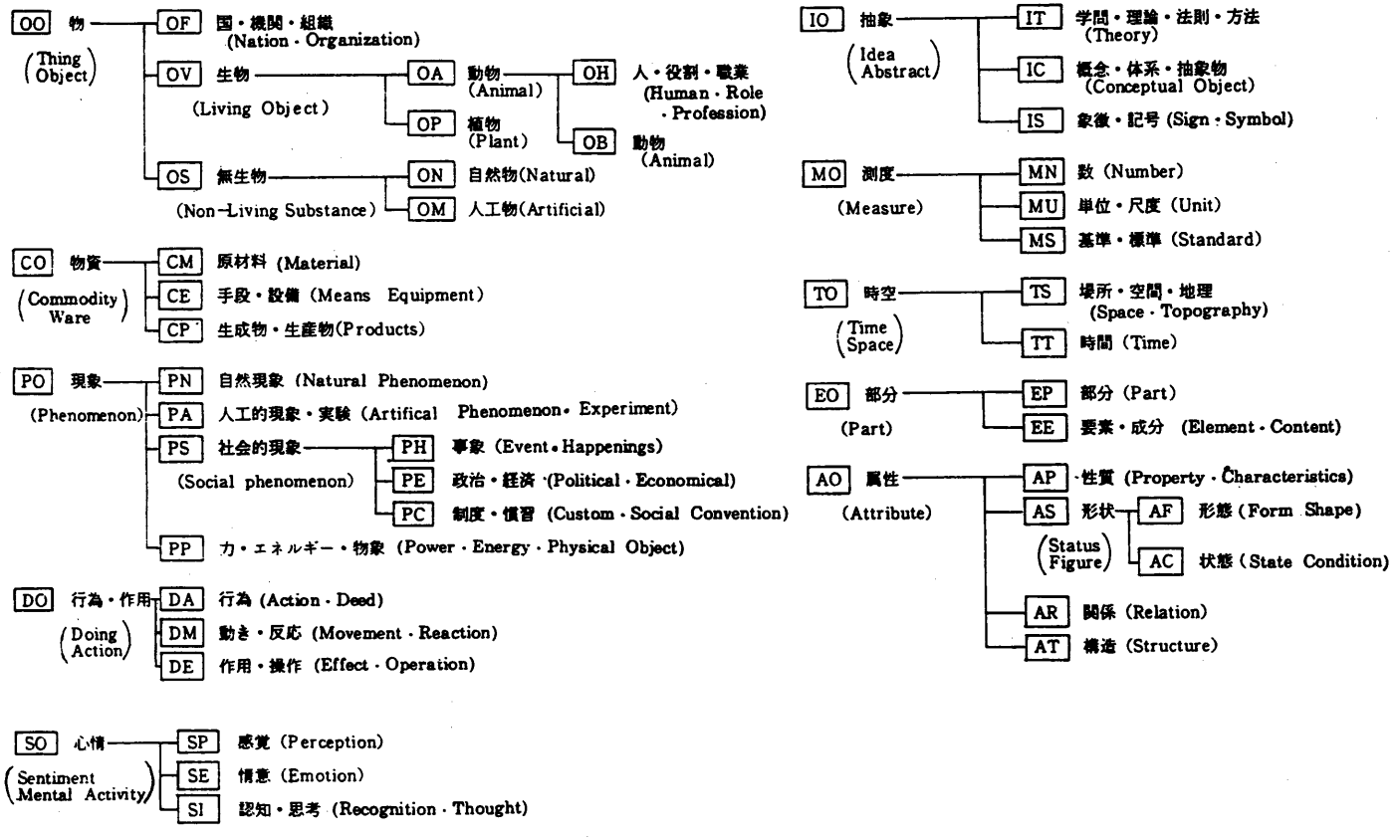
第3表　日本語品詞一覧

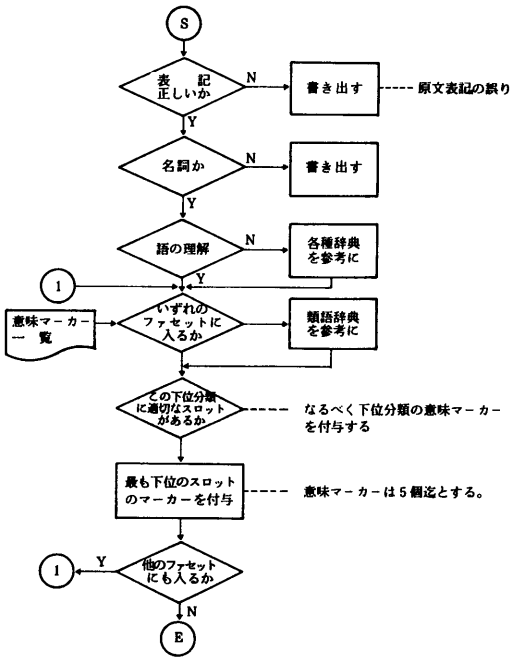
品詞	細分類品詞	記号	用例
名詞(M)	固有名詞	MKY	(組織名，人名，地名等)
	普通名詞	MFT	(例) 自動車，山，構造，深層，緑
	動作名詞（サ変）	MSA	「名詞＋する」（例）概説，利用，運動
	動作名詞（その他）	MST	「連用形動名詞」（例）ずれ，ゆれ，ふれ
	副詞的名詞	MFK	「副詞としても用いられる語」
			(例) 将来，昨年，従来
	格助詞的名詞	MKJ	(例) 中，内，前，間，側
	接続助詞的名詞	MSJ	(例) ため，ところ，とき，場合，際
	補文標識	MHB	(例) こと，もの
代名詞(P)	疑問代名詞	PGM	(人，物，場所)（例）だれ，どれ，どこ
	人称代名詞	PNS	(人)（例）彼，彼女，わたし
	指示代名詞	PSJ	(物，場所)（例）これ，あれ，それ
数詞(N)	数	NNN	(例) 1，2，三，四，百，千，万
	数量詞	NSR	(例) 5㎝，10㎏，10cc，100個所
	冠数詞	NKS	(例) 第，約，昭和
	助数詞	NJS	(例) 回，件，階，語，章，部
接辞(Z)	接頭語	ZTG	(例) 諸，各，全，高，低
	接尾語	ZBG	(例) 別，上，中，前
	接頭辞	ZTJ	(例) 不，非，反
	接尾辞	ZBJ	(例) 的，性，化
副詞(F)	情況副詞	FJK	(文修飾，動詞修飾)
			(例) 結局，極力，ここまで
	程度副詞	FTD	(尺度，極度)（例）非常に，たいへん
	陳述副詞	FCJ	(仮定，疑問，打消，願望，比況)
			(例) もし，いつ，必ずしも
	数量副詞	FSR	(例) たくさん，少し，おのおの
連体詞(R)	指示連体詞	RSJ	(例) この，その，あの
	疑問連体詞	RGM	(例) どの，どのような
	限定連体詞	RGT	(例) ある，さる，あらゆる
	形容詞的連体詞	RKY	(例) 大きな，小さな，少しの

品詞	細分類品詞	記号	用例
接続詞(S)	文接続詞	SBN	(例) したがって，しかし，ただし
	句接続詞	SKU	(例) または，もしくは，そして
動詞(D)	動詞	DSI	(全ての動詞)
助動詞(A)	法助動詞	AHO	(否定，必要性，勧奨，許可…)
	相助動詞	ASO	(相動詞を含む)
	態助動詞	ATI	(受身，使役)
	分裂構文	ABK	(目的，理由)
	その他の助動詞	AST	(補助用言的助動詞を含む)
助詞(B)	格助詞	BKK	(例) が，は，に，と，で，から，より
	接続助詞	BSZ	(例) ば，と，が，のに，ので，から
	副助詞	BFK	(例) も，こそ，さえ，しかし，のみ，だけ
	並列助詞	BHR	(例) と，や，か
	終助詞	BSY	(例) ね，さ，か，よ，わ
	準体助詞	BJT	(例) の，か，がどうか，か否か
形容詞(K)	情意形容詞	KJI	(「語幹＋がる」で動詞になる)
			(例) うれしい，悲しい
	性質・情態形容詞	KSJ	(物，属性，動作を形容する)
			(例) 固い，著しい，速い
	関係形容詞	KKK	(「こと」，「もの」の間の関係を示す)
			(例) 遠い，近い
形容動詞(L)	性質・状態形容動詞	LSJ	(物，属性，動作を形容する)
			(例) 失礼，十分，容易
	関係形容動詞	LKK	(「こと」，「もの」の間の関係を示す)
			(例) 同じ，別

注）形態素解析の結果表示する品詞として以下のものがある。
体述的形容動詞｜LTZ｜「名詞＋だ，である」
(例) JICSTだ，科学技術庁だ

第4表　名詞意味マーカー一覧





第1図　名詞意味マーカー付与作業の流れ

4．体言辞書の作成作業

4.1　名詞辞書

JICST文献速報の電気工学編1982年版6号中の600抄録（2439文）を対象として名詞の辞書データの作成を行った。今後大量の名詞を収集するのに備え，できる限り作業者に負担を与えないことを考慮して，KWICリストを活用することにした。名詞の抽出は次の基準を設けて実施した。

① 基本的には長単位で切り出す。
② 接辞・接語のついている名詞については，これを分離したものも別途抽出する。
③ 語基に分解できるものは，作業者がわかる範囲で別見出しとして抽出する。

上記の基準で抽出した名詞の数は，重複も含めて約12000語になった。

4.2　名詞以外の体言辞書

副詞，連体詞，接続詞については，三省堂の新明解国語辞典より抽出した。

4.3　作業上の問題点

① 名詞単位の認定の困難性
② 品詞細分類の付与の困難性
③ 意味マーカーの付与の困難性
④ 対訳語の付与の困難性
⑤ 副詞の深層情報の抽出の困難性

上記②～④の解消には，作業者の専門用語の意味理解が必要となる。

5．おわりに

実験用の対象文から抽出した体言関係の辞書データについては，単語認定の誤り，品詞細分類及び意味マーカー付与の作業者によるバラツキの見直しなどの修正を現在行なっている。この辞書データに関する情報を，形態素解析，構文解析，トランスファー，生成との関連の中でフィードバックさせていき，問題点を整理していきたい。

このように高度に知的な判断を要し，大量のデータを扱う辞書の作業をいかに推進していくか，いかに機械化していくかが実用化へ向けての要検討課題である。

謝辞

最後に，本研究の実施に当って多くの御指導と御助言を戴いた辞書ワーキンググループの諸先生方，煩雑な言語データの整理などを終始手伝って下さった房珠子嬢ならびにJICSTの各部室の方々に深く感謝いたします。

参考文献

[1] 「辞典事典総合目録」，出版ニュース社，1982
[2] 長尾：科学技術庁機械翻訳プロジェクトの概要（本研究会資料）
[3] 坂本：格構造を中心とした用言と付属語辞書（本研究資料）