自然語キーワード抽出規準の設定とその規準 による主題抽出実験およびその評価

An Experimental Study on Subject Expression by Natural Language Keywords in Japanese

広 井 聡*

奈良県立医科大学附属図書館

1.	序章
2.	自然語キーワードの抽出規準
2.1	カテゴリーの決定
2.2	分析枠の設定
2.3	抽出規準
3.	当抽出規準による主題分析の実験とその評価
3.1	実験データ
3.2	実験方法
3.3	実験結果
3.4	実験結果の分析
3.4.1	カテゴリー分布の分析
3.4.2	特集号タイトルの信頼性
3.4.3	抽出できないタイトルおよび用語
4.	結論

1. 序 章

一般に、文献の内容に着目した場合、文献の整理とは文献内容を主題分析し、文献内容を表わす重要な主題を抽出、付加し、それを一定の基準に基づく索引語という言語記号に変換し、所定の構成のもとにファイルすることであると言われる。又、文献の検索とは質問内容の主題を分析し、質問内容を表わす重要語を選択・付与し、それを索引語と同一の基準に基づく検索語という言語記号に変換し、索引ファイルの中から索引語と検索語をマ

ッチングさせ、必要文献を入手することであると言われる。従って、文献の整理、検索では、いかにして適格に索引語、検索語を決定するかがキーポイントとなる。

索引(主題)作業を主題分析、主題抽出、索引語決定の順に一般的に述べると次のようになる。

文献のマニュアルな主題分析法として、現在、最も普遍化されている方法は領域ごとにカテゴリーを設定し、そのカテゴリーによって主題を分析する方法がとられている。例えば、次の基本カテゴリーがあれば、ほぼ全領域の主題を分析できると言われる。(1)構造体、(2)部分、(3)物質、(4)構造、(5)運動、(6)作用、(7)性質、(8)操作、(9)空間、(10)時間¹⁾。医学分野では、(1)器官・器具、(2)症状・現象、(3)用途・性質、(4)対象・能因、(5)操作・処置²⁾等のカテゴリー群が考えられる。

次に主題抽出であるが、これは主題分析によって得られた主題から文献検索に重要と考えられる主題を抽出する。医学分野では、前記した(1)から(5)の基本カテゴリーのなかでは、(1)器官・器具、(2)症状・現象、(4)対象・能因、(5)操作・処置等のカテゴリーに属する主題が重要な主題であると言える。

最後に索引語の決定であるが、これには次の方法がある。(1)抽出された主題は自然語で表現されているが、これをそのままキーワードとする。(2)前記の自然語をこれに対応するシソーラス中のディスクリプターに、1対1、又は1対多で置換する。(3)自然語群に対応するディスクリプターに置換する。

* Satoshi HIROI: Nara Medical University Library

索引語を決定するには、前述したように主題分析、主題抽出が先に行なわれねばならない。原文献を主題分析することによって得られるものは、内容の詳しさの順に抄録、標題、索引となる。適切に索引語を決定するには、適切に標題を決定することが前提となり、適切に標題を決定するには、適切に原文献が主題分析され、抄録が作成されていることが前提となる。すなわち、索引語決定のための主題分析は原文献から始めるのが理想的な形態である。しかしながら実際問題として、どの程度詳しく主題分析するかはスタッフの規模、および主題に対する能力によって決定されるであろう。

Weil⁸⁾は原文献、抄録、標題、索引の間の情報伝達量の関係を述べたあと、それぞれの役割の違いを次のように述べている。「原文献や抄録等の文献を読むことによって、実際に情報を伝達することと、ある文献がある主題に関係があると認めることは別のことである。後者の目的には、標題は抄録と同程度に十分な役割を果たしている。」又、Luhn⁹⁾のKWICの仮説、すなわち、「論文標題は最も忠実に主題を反映したものである」という仮説はKraft⁵⁾やMontgomery⁶⁾によって実証されている。上記のことから標題もしくは抄録を主題分析し、索引語を決定しても原文献の主題を表示できると考えてもよいであろう。

以上、索引語決定の重要性、主題分析、主題抽出、索引語の選択、主題分析の範囲について述べてきたが、ここでは、索引語を自然語キーワードとし、主題分析の範囲を和雑誌特集号の標題とし、対象領域を医学に限定した場合、主題分析、主題抽出の規準、すなわち自然語キーワード抽出の規準となる基本カテゴリーは、いかなるものを設定すべきか、又、今回設定された基本カテゴリーによって抽出された主題はいかなる特質をもっているか、更に雑誌特集記事のタイトルは原文献の主題を適切に表現しているかどうか、この抽出規準ではどのような主題が抽出もれとなるか等の点について述べたい。

2. 自然語キーワードの抽出規準

2.1 カテゴリーの決定

文献のマニュアルな主題分析法として、現在、最も普遍化されている方法としては、領域ごとにカテゴリーを設定し、そのカテゴリーによって主題を分析していく方法がとられていることは、前述したが、それでは医学の場合どのようなカテゴリーを、どのような方法で設定すればよいのか。

カテゴリーの設定には、演繹的な方法と帰納的な方法がある⁷⁾が、厳密に演繹的方法と帰納的方法を区別することは、総ての学問の方法論と同様に無理があるであろう。実際的な方法としては、演繹的方法と帰納的方法を併用することが妥当である。

坂本徹朗氏によればカテゴリーの設定は次のような演繹的-帰納的な順序で行なわれる⁸⁾。

(1) 対象物とする主題分野を決める。たとえば、物理学、化学、医学など。

(2) その主題構造分析を行ない、カテゴリーを見つける。

(i) その分野に関する既存の分類表、索引、抄録誌、ハンドブックなどを集める。

(ii) 既設の基本カテゴリーを参照しながら、(i)に現われる主題語をまとめ変えていく。

(iii) 主題語グループが互いに分離よくまとまったところまでまとめ変えを終え、そのグループに適当な名をつけカテゴリーとする。

(3) その分野の論文で種々タイプの異ったものを集める。

(4) それを先のカテゴリーで分析してみて、拾いもれ、不均衡がないかテストする。

(5) 分析もれがある場合には、(2)の(i)をその論文に置換え(2)~(4)を繰り返す。

(6) カテゴリーが安定したら、それに適当な順序を決める。たとえば、“原料→生産物”の順序とか、“具体物→抽象概念”の順序などである。

医学領域の主題分析を行なうために必要なカテゴリーを決定する際に、次のような医療のプロセスが参考になる。すなわち、いかなる生体のいかなる構成部分あるいは物質が、いかなる原因により、いかなる状態になり、どのような疾病、傷害、病症をおこしたか。その予防、基礎的研究、診断、治療、予後は、どのような生体、生体構成成分、もしくは物質を用い、どのような条件のもとで、どのような措置を使い、どのような方法で操作、処置したかというようなプロセスである。

このようなプロセスを分析してみると次の2つのカテゴリーのグループができる。

(1) 予防、基礎的研究、診断、治療、予後

予防基礎的研究：解剖、発生、生理、生化、薬理、病理、病因(原)、免疫

診断：診査法／視診、触診、打診、聴診

：検査法／化学的検査法、糞便検査法、尿検査法、

血液検査法, 体温検査法, 反射検査法, 細菌学
の免疫学的検査法, 組織学的検査法

治療: 対症療法, 技術療法, 化学療法, 血清療法, 理
学的療法, 物理療法, 食餌療法, 精神療法
: 看護

予後

(2) 疾病, 傷害, 症状, 生体, 生体構成部分, 物質,
性質(程度, 形態, 運動, 作用, 機能, 状態, 現象, 操
作, 方法, 処置), 操置, 設備, 材料, 時間, 空間, 研
究・診療の領域

これを更にグループ分けすると次のようになる。

(1) 予防, 基礎的研究, 診断, 治療, 予後

予防

基礎的研究: 解剖, 発生, 生理, 生化, 薬理, 病理,
病因(原), 免疫

診断: 診査法, 検査法

治療: 療法, 看護

予後

(2) 疾病, 傷害, 症状, 生体, 生体構成部分, 物質,
性質(形態, 程度), 運動(作用, 機能), 状態(現象),
操作(方法, 処置), 操置(設備, 材料), 時間, 空間,
研究(診療)領域。

カテゴリーを決定する際, 以上の処置は, 暫定的な処
置であって, 前述したように, これらのカテゴリーは,
サンプルによって実験され, 拾いもれ, 不均衡等がテス
トされる。もし分析もれ, 不均等がある場合には, 坂本
氏が述べているように⁹⁾, (2)~(4)の作業を繰り返さ
ねばならない。

一応, 暫定的にカテゴリーが決定したとして, 次に,
このようなカテゴリーをどのように使用して主題を分析
していけばよいのか, その方法を検討してみたい。

2.2 分析枠の設定

表 1 主題語抽出の分析枠

研究診療領域	機能・運動・作用	性 質（程度）	生物	生体構成部分	生	症	傷	疾	予	治療		診断		基礎研究								
										看	療	検査法	診査法	免	病	病	病	葉	生	生	發	解
免疫学・小児科・外科	反射運動・呼吸運動	先天性	ホルモン	腎臓・上肢	小児・犬・球菌	貧血	打撲・捻挫・骨折	糖尿病・腎臓病	リハビリテーション	小児看護	輸血・注射・精神療法	免疫学的検査法	視診・触診・打診・問診	免疫学的検査法	スピロヘータ・球菌	貧血・先天性奇形	向精神剤	蛋白質・脂質・ホルモン	呼吸・排泄・妊娠・代謝	胚葉發生	気管・咽頭・腎臓・上肢	予防注射

主題分析の実際的な作業は、これらのカテゴリーをマトリックス化し、分析枠を設定して、文献中に枠内カテゴリーに相当する主題語があれば、残らず抽出し、該当枠中に列記する。例えば、上記のカテゴリーの場合、表1のように枠組みし、主題語を抽出する。

この場合、一見してわかることは同一の主題語（注射、腎臓、呼吸、貧血、小児、ホルモン等）がいくつものカテゴリー枠に、二重、三重、四重に出てくることである。これは各カテゴリー間の関係が明確になっていなかったからである。各カテゴリー間の重複をなくするには、一次元の分析枠では不可能である。腎臓病が予防に関係があるのか、腎臓病の生理に関係するのか、腎臓病の治療

に関係があるのかを一目瞭然と区別するには次のような二次元的な分析枠を設定しなければならない。例えば、

- (i)：腎臓病の尿検査法と食餌療法
- (ii)：ホルモン代謝の異常と生理

は表2のようになる。

ここで、腎臓、尿が縦のカテゴリーで、診断および治療の両カテゴリーに重複して出てくるのはそれらが、診断および治療に関係している文献内容を持っているので当然であると判断できるが、横のカテゴリーで、疾病、生体構成部分、物質、操作（処置、方法）の各カテゴリーに重複して出てくるのは、斎藤氏が言うように漢字は、その合成によって非常に造語能力が大きく、日本文

表 2

		疾 病	傷 害	症 状	生 体	生体構成部分	物 質	～	作用・運動・機能	状態現象	操 作	処置・方法	～
予 防								～					～
基礎的 研究	解 剖							～					～
	発 生							～					～
	生 理						ホルモン	～	代謝	異常			～
	生 化							～					～
	薬 理							～					～
	病 理							～					～
	病 因							～					～
	(原)							～					～
診 断	免 疫							～					～
	診 査							～					～
	検 査	腎臓病				尿・腎臓		～			尿検査法		～
治 療	療 法	腎臓病				腎臓・尿		～			食餌療法		～
	看 護							～					～
予 後								～					～

は単語の区分が不明確である¹⁰⁾ ことによると考えられる。

機械検索(コンピュータ、カードセレクター等)の場合、腎臓、尿とインプットするか、腎臓病、尿検査法とインプットするかで、インプット、ファイル、アウトプットの構成に非常に大きな違いが生じ、重大な問題が起きてくるのであるが、これについては次の機会に述べることにしたい。

ここでは、主題分析が当面の課題であるので、このような合成語を主題分析する際にどのような問題が生じるか述べることにしよう。たとえば、「高血圧」を一語としてとらえた場合、横のカテゴリーでは、「症状」もしくは、「運動・作用・機能」あるいは「状態・現象」のいずれかに、縦のカテゴリーによって分析されるが、これを「高」と「血圧」の二語よりなる合成語としてとらえると「高」は「性質・形態・程度」を表わすカテゴリーに分析され、「血圧」は「運動・作用・機能」、もしくは「状態・現象」のカテゴリーに分析され、「症状」というカテゴリーには全く属さなくなる。

これは、カテゴリーの設定にも問題がある。状態、現象は症状の上位概念であると考えられ、同一のカテゴリーとして、扱うべきであると考えられるからである。ここで、これを2つのカテゴリーとして独立した位置を与えたのは、医学研究の扱う範囲が病症だけとは限らないからであり、にもかかわらず、特に病的状態を多く扱うからである。

更に、これはカテゴリー設定以前の問題として言語の意味の持つ本質的な問題にかかわってくる。そもそも、言語の意味が場的な構造によって規定されてくるものであるとすれば、索引、又は検索のために言語を、バラバラに切離し、抽象化し、カテゴリー化しようとすることに無理が生じるのである。この意味において、Luhn の KWIC-Index は成功した例であろう。しかしながら、我々が情報の整理、蓄積、検索の仕事に携わっている以上、言語情報を最も効率よく伝達するためには、この抽象化は避けられない宿命である。ここで、現実的になし得ることは、可能な限り言語情報の持つ意味を正確に伝達するために、言語を索引語化、検索語化するには、それを行なう前に、あらかじめ、約束ごとをもうけることである。複合主題の主題分析は、この二次元的機能を持つカテゴリー表によって、一応、分析できると思われるので、次に、実際に和雑誌の特集号タイトルを主題分析するにあたって、細かい規準を設定してみたい。

2.3 抽出規準

序章で述べたように、ここでは、主題分析によって抽出された自然語をそのまま索引語とするのであるから、カテゴリー表そのものが、そのまま最も重要な抽出規準となるのであるが、「高血圧」のような合成語をどのように処理するかという問題も含め、細かい規準を設定してみたい。

(1) 日本語および日本語化された外来語に限る。それ以外の外国語は日本語に翻訳する。

(2) 医学関連領域の和雑誌特集号の特集タイトルより助詞、助動詞、接続詞、代名詞、形容詞、動詞、副詞を除いた重要語を抽出する。

(3) 合成語は意味を全く失なわない限り、できるだけ各カテゴリーに細分する。

(i) 各カテゴリーの見出し語、もしくは見出し語に類似した主語が合成語となっている場合は抽出しない。

例：食餌療法—これは「食餌」のみを抽出する。

(ii) (3)—(i)によって、あるカテゴリーに属する内容を持っていた合成語がそのカテゴリーに入れるべき用語を失った場合は「*」印を分析枠につける。

例：尿検査法—“尿”は生体構成部分に属し、“検査”、“法”は(3)—(i)の適用によって抽出しないので、操作、処置、方法のカテゴリーに入れるべき用語がなくなった場合。

(iii) 合成語を分析した結果、全く意味を失う場合は分析しない。

例：出血—“出”を運動に、“血”を物質とはしない。

例：輸血—“輸”を運動に、“血”を物質とはしない。

前者の場合は症状に、後者の場合は操作、方法、処置に入れる。

(iv) 性質、程度を現わす接尾語、“～性”、“～化”、“～様”、“～状”、“～的”等と結びつく用語が他のカテゴリーに属する場合は、他のカテゴリーに入れる。

例：絨毛性—これは“絨毛”を生体構成部分に入れ、“性”は抽出しない。

例：化学的—これは“化学”を研究、診療領域に入れ、“的”は抽出しない。

(v) 性質、程度を表わす接頭語と結びつく用語が

表 3 カテゴリー別抽出用語

疾 病		症 状		生 体		生体構成部分	
膠原(病)	14	病 変	5	胎 児	11	心	22
リウマチ	12	炎 症	5	老年者	9	筋	21
狭心(症)	8	麻 痺	4	妊 婦	7	血 液	21
結 核	6	化 膿	4	患 者	7	動 脈	20
てんかん	5	失 調	4	自 己	7	消化管	19
痛 風	5	腫 瘤	4	医 生	5	精 神	17
感 冒	5	下 痢	3	老 年	4	心 臓	17
類肉腫(症)	3	めまい	3	生 物	3	脈 管	17
狼 瘡	3	血 栓	3	高令者	3	甲状腺	16
か ぜ	2	流 産	3	医 家	3	大 腸	16
イレウス	2	頭 痛	3	成 人	3	脾	15
赤 痢	1	愁 訴	2	医 師	2	腸	15
疫 痢	1	便 秘	2	保健婦	2	肝 臓	14
痔	1	痔 瘻	2	未熟児	2	骨	14
躁うつ(病)	1	湿 疹	2	真 菌	2	膜	13
血友(病)	1	DOWN症侯群	2	汚染者	1	頭 部	12
梅 毒	1	母 斑	2	者	1	指 腸	12
油 症	1	けいれん	2	看護婦	1	リンパ	11
後遺症	1	発 熱	2	学 童	1	髓	10
イタイイタイ病	1	変 異	2	原生動物	1	染色体	10
		肥 満	1	乳 児	1	臓 器	10
		るい瘦	1	高 年 者	1	子 宮	10
		肥 胖	1	薬剤師	1	担道系	10
		疲 労	1	ホームヘルパー	1	腹 部	9
		痴 呆	1	動 物	1	消化器	8
		ふ る ぜ	1	児	1	心 身	8
		胸 痛	1	カ ビ	1	腎 臓	8
		奇 形	1	球 菌	1	皮 膚	8
		癌 化	1	桿 菌	1	脾 臓	7
		こんすい	1	救 患	1	腹	7
		カルチノイド	1	結核菌	1	腰	7
		うっ血	1	宿 主	1	絨 毛	6
		びらん	1	細 菌	1	胆	6
		流 血	1	酵 母	1	気管支	6
		動 悸	1	マイクロバイ	1	食 道	6
		忘 相	1	病原体	1	組 織	6
		疼 痛	1			尿 路	5
		むくみ	1			胎 盤	5
		痔 核	1			循環器	5
		気 腫	1			胸 腺	5
		死	1			下垂体	5
		死 亡	1			股	5
		仮 死	1			肢	5
						肩	5
						間 脳	5
						担 道	4
						頭 蓋	4
						頸 部	4
						泌尿器	4
						膣	4

傷 害	
障 害	47
外 傷	35
骨 折	11
脱臼	9
損 傷	4
創 傷	2
欠 損	2
捻 挫	1

症 状	
癌	88
異 常	43
炎	43
腫 瘍	40
潰 瘍	21
出 血	20
腫 脹	19
黄 疸	15
血 貧	15
アレルギー	10
中 毒	10
腰 痛	9
熱 卒	7
閉 塞	7
腹 痛	6

生 体	
小 児	42
新生児	33
ビールス	21
老 人	17

生体構成部分	
神 経	48
胃	45
細 胞	43
血 管	39
脳	36
尿	36
肝	36
腎	34
関 節	28
肺	24
血	24

生体構成部分		生体構成部分		生体構成部分		物 質	
骨 盤	4	漿 水	1	葉緑体	1	P C B	2
呼吸器	4	腔 部	1	気 胸	1	薬 剤	2
脂 血	4	股 部	1	動脈壁	1	毒	2
手	4	関節部	1	皮	1	強心薬	2
網内系	4	憩 室	1	腕	1	カドミウム	2
胆 石	4	原形質	1	神経叢	1	ヨーソ	1
羊 水	4	芽細胞	1	胆のう	1	ISOTOPE	1
小 腸	4	平滑筋	1	耳	1	ワクチン	1
顔 面	3	尿 酸	1	鼻	1	農 薬	1
神経系	3	循環系	1	膝	1	ピ ル	1
腺	3	下垂体系	1	ミトコンドリア	1	ロイトマイシン	1
全 身	3	顎	1	色 素	1	ゴナドトロピン	1
体 液	3	滑 面	1	膀 胱	1	脂生物質	1
子宮頸	3	小胞体	1	身 体	1	蛋白質	1
直 腸	3	関 門	1	基靱体	1	酸化系物質	1
卵 管	3	幹細胞	1	表 面	1	原 子	1
腹 膜	3	縦隔膜	1	左心室	1	窒素化合物	1
弁 膜	3	坑 原	1	右心室	1	有毒物	1
眼	3	白血球	1	局 所	1	元 素	1
脾 臓	3	結合組織	1	胆色素	1	分 子	1
リンパ球	3	義 肢	1	鼠経部	1	アミノ酸	1
血 漿	3	頸 管	1	心齋部	1	アミン	1
線 維	3	リンパ球	1	器 管	1	配位子	1
皮 質	2	系	1			カリウム	1
単核球	2	乳 腺	1	物 質		カルシウム	1
足	2	頸 髓	1	糖	33	鉄	1
血 清	2	視 路	1	薬 物	20	タバコ	1
背 髓	2	肛 門	1	ホルモン	13	ステロイド系	1
頸 器	2	臍	1	酵 素	11	リボゾーム	1
胸部	2	胞 子	1	R I	11	ポリシン	1
結晶体	2	補 体	1	剤	10	ジギタリス	1
背 椎	2	母 乳	1	脂 質	9	クロマチン	1
頸 椎	2	乳 房	1	R N A	8	ポリペプチッド	1
頭 椎	2	血 球	1	食 品	8	インシュリン	1
髓 液	2	附属器	1	食 品	1	水 銀	1
卵 巢	2	精 子	1	蛋 白	8	メチール	1
腹 水	2	性 腺	1	ガ ス	7		
粘 膜	2	喉 頭	1	ステロイド	4		
腹 壁	2	虫 垂	1	核 酸	3		
気 道	2	へ ぞ	1	ペプチッド	3	性質・形態・程度	
体 液	2	卵細胞	1	グロブリン	3	高	26
腿	2	松果体	1	ヘテロ環	3	悪 性	23
坑 体	2	胞 体	1	食 餌	3	不 全	19
糸球体	2	配糖体	1	電解質	2	副	10
直 腸	2	核	1	エストロゲン	2	難	6
上 皮	1	胃 液	1	L H	2	特殊性	5
脳 系	1	走血球	1	コレステロール	2	良 性	4
門 脈	1	塩基球	1	医薬品	2	活 性	4
液 体	1	酸 球	1	ビタミン	2	特異性	4
乳頭部	1	乳 汁	1	糖 質	2	硬	3
肘	1	骨 格	1	金属錯体	2	單純性	3
変異株	1	核小体	1	製 剤	2	重	3
		葉緑素	1	アルコール	2	多発性	3

性質・形態・程度		運動・作用・機能		運動・作用・機能		状態・現象	
自然	2	分泌	39	再生	2	公害	8
微小	2	代謝	37	言語	2	老化	6
不明	2	感染	27	交通	2	事故	3
低低	2	硬化	16	維持	2	生命	3
低	2	妊娠	16	着床	1	危容	2
難治性	2	シヨック	14	生殖	1	線容	2
簡易	2	循環	13	軟化	1	変形	2
原発性	2	呼吸	13	変化	1	血流	2
天然性	1	分娩	10	感受	1	健康	2
清潔	1	吸收	10	抗疑	1	複合	2
不均等	1	光合成	9	酸化	1	死	2
優性	1	抗生	9	吸光	1	安全	2
好多	1	消化	7	制御	1	情報	2
効果的	1	適応	7	流動	1	行効	2
超微細	1	自律	6	収縮	1	血藥	1
從属的	1	脱合	6	平衡	1	無産	1
細	1	伴妊	6	刺激	1	安性	1
潜在的	1	不妊	6	伝達	1	成分	1
耐性	1	反応	5	復帰	1	立布	1
微量	1	放射	5	侵襲	1	相関	1
巨大	1	線自	4	蘇防	1	緩和	1
複雑	1	透過	4	減数	1	到達	1
親	1	遺伝	4	發育	1	発見	1
不良性	1	発癌	4	受精	1	生	1
非	1	随伴	3	制酸	1	加	1
習慣性	1	随伴	3	視覚	1	リズム	1
巨	1	出産	3	抗下	1	浸潤	1
依存性	1	凝固	3	抗うつ	1	限界	1
否定	1	進化	3	発達	1	滯害	1
汎	1	排尿	3	灌流	1	遲害	1
悪性	1	汚染	3	自動化	1	災再	1
不整	1	分裂	3	促進	1	超過	1
微	1	抑制	3	集合	1	迷切	1
易	1	生活	3	寄生	1	寿命	1
大量	1	栄養	3	食思	1	騒音	1
硬度	1	保存	3	閉鎖	1	心理	1
偶発性	1	進出	3	排水	1	成果	1
経口性	1	放出	3	意行	1		
月状	1	亢進	2	行動	1		
ブドウ状	1	転位	2	温造	1		
奇形	1	結合	2	造血	1		
空洞	1	同化	2	生動	1		
体質	1	隆起	2				
奇胎	1	形成	2				
有毛	1	支配	2				
		分化	2				
		癌化	2				
		抑制	2				
		抗菌	2				
		調節	2				
		調抗	2				
		成長	2				

操作・方法・処置	
管理	27
輸液	18
応用	12
輸血	12
人工	11
切除	7
麻醉	7
移植	6
子防接種	5
帝王切開	5
造影	4
中絶	3

操作・方法・処置		操 置・材 料		空 間		研究・診療領域	
遠 隔	3	心電図	11	前 後	1	腫瘍診断学	1
指 導	3	内視鏡	8	上 部	1	形態学	1
避 妊	3	電子計算機	5	広 範	1	実地医学	1
撮 影	2	CHROMATO GRAPY	4	範 囲	1	人類遺理学	1
置 換	2	電子顕微鏡	4	下	1	基礎医学	1
剥 離	2	シンチグラム	2	周 辺	1	境 域	1
逢 合	2	X線像	2	社 会	1	VIRUS学	1
投 与	2	筋電図	1	侵入路	1	思春期小児科学	1
アプローチ	2	機 器	1			老人医学	1
応 急	2	電顕図譜	1			小児歯科	1
注 入	1	機 械	1			大腸外科	1
固 定	1	アンジオグラム	1			美容外科	1
遮 断	1	心音図	1			低温医学	1
走 査	1	心機図	1			酵素化学	1
圧 迫	1	心エコー図	1			スポーツ医学	1
接 合	1					社会心理学	1
命 名	1					災害医学	1
減 圧	1					消化器外科	1
採 血	1					光化学	1
読 影	1					有機化学	1
併 用	1					病院精神医学	1
吻 合	1					免疫化学	1
分 離	1					産婦人科領域域	1
切 迫	1					分子科学	1
計 画	1					診断学	1
保 健	1					各 科	1
統 計	1					法 律	1
指 導	1					肝臓病学	1
SCREENING	1					症候診断学	1
対 話	1					耳鼻咽喉科	1
注 射	1					血清学	1
透 導	1					精神身体医学	1
給 食	1					神経生物学	1
視 折	1					周産期免疫学	1
全 剔	1					免疫血液学	1
保 護	1					有機化学	1
選 択	1					発達生理学	1
隆 圧	1					低温生物学	1
対 比	1					天然物化学	1
作 業	1					老人外科	1
調 圧	1					神経内分泌学	1
鎮 痛	1					薬 学	1
知 見	1					リハビリテーション医学	1
点 検	1					細胞病理学	1
切 開	1					電子顕微鏡学	1
比 較	1					老人症候学	1
根 治	1					看護学	1
常 用	1					遺伝子学	1
						心臓病学	1
						M E	1
						血行化学	1

時 間	
慢 性	29
先 天	18
後 期	11
急 性	10
時	8
白	8
緊 急	7
前 期	5
時 代	2
末 期	2
中	2
突 然	2
初 期	2
夏 季	1
青春期	1
老年期	1
新 鮮	1
夏	1
時 期	1
高 速	1
日 常	1

空 間	
内	44
外	36
末 梢	9
日 本	4
実 地	3
環 境	3
上	2
底	2
高 位	1
基 底	1
地 域	1

研究・診療・領域	
臨 床	134
外 科	52
化 学	28
産婦人科	21
基 礎	19
内 科	13
医 学	8
産 科	7
免疫学	7
婦人科	5
生化学	5
臨床生化学	5
小児科	4
周 辺	4
内分泌学	4
眼 科	3
疫 学	3
血液学	3
腎臓病学	3
生理学	3
錯体化学	3
形成外科	3
循環器学	3
精神医学	3
生物学	3
M E	2
臨床病理学	2
動力学	2
結核病学	2
臨床医学	2
核医学	2
病態生理学	2
実 際	2
公衆衛生	2
境界領域	2
精神科	2
神経(学)	2
小児放射線学	2
臨床免疫学	2
周産期医学	2
皮膚科学	1
腎臓形態学	1

他のカテゴリーに属する場合は他のカテゴリーに入れる。

例：高血圧―“高”は性質，程度のカテゴリーに，“血圧”は作用，機能か，状態，現象のカテゴリーに入れる。

(4) 数字は全く抽出しない。

(5) 合成語を分析した結果，同一のカテゴリー内に入る場合は，列記する。

例：腰椎椎間板―腰椎，椎間板

(6) 学問領域を表わす接尾語“学”，“科”がつく合成語はすべて学問領域のカテゴリーに入れる。

(7) (3)～(4)の場合で，“～性”，“～様”，“～状”“～的”等の接尾語と合成してできる用語が他のカテゴリーに入らない場合は，接尾語と共にすべて抽出する。

3. 当抽出規準による主題分析の実験とその評価

3.1 実験データ

実験データは 奈良県立医科大学附属図書館で昭和43年より蓄積していた医学関連領域の特集記事目録より，和雑誌特殊記事の目録，1890件を対象とした。

3.2 実験方法

2.3の抽出規準により，表2のカテゴリー表を用いて，特集タイトルより主題を抽出し，表に列記していった。実施するに際し，(2)～(3)の(i)，および(2)～(3)の(iv)，(2)～(4)に属する用語と時間のカテゴリーに属する合成語に多く現われる～期は，非抽出用語リストを別に作成し，集計した。

3.3 実験結果

主題分析の結果，各カテゴリーに表3のような用語が抽出された。右側の数字は出現頻度を表わす。

また，各カテゴリーの非抽出語の集計結果は次のようになった。()内の数字は出現頻度を表わす。

疾病：疾病(7)，疾患(186)，～病(70)，～症(117)

傷害：なし

症状：症状(2)，病態(43)，症候群(11)，症候(7)，症例(5)

生体：生体(8)

生体構成部分：臓(93)，器(35)，管(69)，～系(19)，～部(33)，～組織(3)

物質：物質(12)

性質・形態・程度：～性(177)，～的(33)，～状(2)，～面(1)，形態(4)，構造(9)

運動・作用・機能：運動(8)，作用(12)，機能(39)，

機序(6)

状態・現象：現象(2)

操作・方法・処置：処置(14)，処理(14)

操置・設備・材料：機器(1)，機械(1)

時間：～期(41)

空間：なし

研究・診療領域：なし

予防：予防(7)

解剖：なし

発生：発生(8)

生理：生理(43)

生化：なし

薬理：薬理(4)

病理：病理(16)

病因・病原：病因(5)，原因(4)，成因(5)

免疫：免疫(53)

診断・診査：診断(88)，診療(34)，鑑別(8)，～診(5)，診断法(5)，内診(3)，初診(3)，視察法(1)，識別(1)，見方(1)，解診(1)

診断・検査：検査法(17)，検査(42)，測定(7)，分析(9)，実験(4)，生検(3)，集積(3)，診査(1)，検診(4)，予知(1)，読み方(1)，テスト(1)，解析(3)，分析法(1)，測定法(3)，実験法(1)，定量法(1)，所見(2)，計測(1)

治療：治療(133)，治療法(5)

療法：療法(67)，診療(34)，医療(18)，処法(1)，療養(1)，とりあつかい方(4)

看護：看護(6)

予後：予後(5)

3.4 実験結果の分析

医学関連領域の和雑誌特集記事タイトル1,890件より抽出された延主題数は，3.3で列記した非抽出語を除いて，3,678語であった。内訳は，疾病(74)，傷害(111)，症状(441)，生体(205)，生体構成部分(1,095)，物質(224)，性質・形態・程度(174)，運動・作用・機能(404)，状態・現象(66)，操作・方法・処置(179)，操置・設備・材料(44)，時間(114)，空間(114)，研究・診療領域(443)であった。カテゴリー別の割合は表4のとおりである。数字はパーセンテージの概算を示す。また，各カテゴリー別用語数を重複してカッコ付き数字で示す。

3.4.1 カテゴリー分布の分析

イ) 横のカテゴリーの分析

横のカテゴリー分布の状況を見ると，最高が生体構成

表 4 カテゴリー別用語数とその割合

カテゴリー	生体構成部分	症 状	研究・診療領域	運動・作用・機能	物 質	生 体	操作・方法・処置	小 計
%	30	12	12	11	6	5.5	5	81.5
用語数	(1,095)	(441)	(433)	(404)	(224)	(205)	(179)	2,981

カテゴリー	性質・形態・程度	空 間	時 間	傷 害	疾 病	状態・現象	操置・設備・材料	大 計
%	5	3	3	3	2	1.5	1	100
用語数	(174)	(114)	(114)	(111)	(74)	(66)	(44)	3,678

部分の1,095語;30%,最低が操置・設備・材料の44語;1%と極端に開きがある。生体構成部分が全体の30%をしめているのは医学の目的が生体の疾病,傷害部分を治療することであることを考えると当然であるが,これはまた内頸動脈静脈叢,頸動脈糸球等,生体構成部分同志の結合が多いこと,さらに他のカテゴリー用語との結合が非常に多く,それらの合成語をできる限り分解した結果であるように思われる。例えば,研究・診療領域のカテゴリー用語とは腎臓病学,運動・作用・機能のカテゴリーとは血液凝固,物質のカテゴリーとは神経下垂体ホルモンというように極めて結合能力が大きい。症状のカテゴリーが次に多いのも医学用語としては当然である。次に研究・診療領域を示す用語が3番目にリストアップされているのは特集号がレビュー誌の役割を果たしているために必然的に概念レベルの高い用語が上位に集ったことによると思われる。疾病のカテゴリーに入る用語が少ないのは疾病と他のカテゴリーの用語,特に生体,生体構成部分を示す用語との結びつきが多く,その場合,3.3の非抽出用語の結果でみたように疾病(7),疾患(186),～病(70),～症(117)という疾病を表わす用語が接尾語として働く場合が多いことによると思われる。それに比べて,傷害を表わす用語が少なかったのは以外であった。この原因は今後,追求していきたいと思っている。状態,現象を表わす用語が最後から2番目にリストアップされているのは,このカテゴリーが,運動・作用・機能のカテゴリーや,症状のカテゴリーとオーバーラップしていることによる。たとえば,疲労,出血,進行,再生等。最後に操作・設備・材料のカテゴリーが1%たらずであったのは,これもまた,特集号のレビュー誌的な性格によると思われる。

総じて概念レベルの高いのは特集号のタイトルとして

は当然であり,利用者層を考えても容易に想像のつくところである。

この分布状況からみると生体構成部分は,細胞,組織,脈管,呼吸器系,消化器系,泌尿器系と,骨格系,神経系,総皮,感覚器系,局所の2つに分ける。運動,作用機能と状態・現象を一つにする。生体に物質を,操作・方法・処置に操置・設備・材料を吸収する。また性質・形態・程度と空間・時間を・さらに疾病と傷害を併合すると,カテゴリー間の分布状況がほぼバランスのとれた状態になる。これをパーセンテージで示すと,上位から順に,15%,15%(共に推定),12%,12.5%,11.5%,6%,11%,5%となる。空間,時間に属する用語の分析はしなかったが,表3を見ればわかるように,空間,時間に属する用語は性質,程度に関連する用語が多い。以上の点からこの新カテゴリー表はある程度,実験に裏づけられた妥当性があると思われる。カテゴリーを決定するには,常に演繹―帰納的な紆余曲折が必要である。次の機会に同種のデータを用いて,このカテゴリー表をテストしてみたい。

ロ) 縦のカテゴリーの分析

3.3の結果をみると診断・治療に関するカテゴリー見出し語およびその類似語が基礎研究の生理,免疫を除くと圧倒的に多いが,(診断―88,診療―34,治療―133,療法―67,に対し,生理―43,免疫―53),このカテゴリー表の横列は,相互に関係が曖昧である。生体の生理が問題なのか,病因として問題になっているのか,横のカテゴリーだけでは,はっきりしない。従って,縦のカテゴリーが,必然的に必要となってくる。ただ,看護は特別にカテゴリーとしてあげる必要はないように思われる。

3.4.2 特集号タイトルの信頼性

表 5 カテゴリー含有数とその割合

カテゴリー		含有数	0	1	2	3	4	5	6	計
横のカテゴリー	件数		0	411	635	559	222	50	13	1,890
	%		0	22	33	29	12	3	1	100
縦のカテゴリー	件数		1,364	408	116	2	0	0	0	1,890
	%		72	21	6	1	0	0	0	100

医学文献のタイトルが文献内容を適格に伝えるには、縦のカテゴリーのいずれか一つ、横のカテゴリーのうち疾病、傷害、症状のいずれか一つ、生体、生体構成部分、物質のいずれか一つ、操作・方法・処置、操置・設備・材料のカテゴリーのいずれか一つをそれぞれ最低限含んでいなければならない。すなわち、このカテゴリー表では、縦のカテゴリーを最低一つ、横のカテゴリーを最低3つ、計4つのカテゴリーを最低限含んでいなければならない。

特集号のタイトルは、特集号という性格、および利用者層の要求内容を考慮しなければならないことはいうまでもないが、ここでは、それらのことを考慮に入れず検討してみたい。

方法としては、この実験データの一件が横のカテゴリーをいくつ含み、縦のカテゴリーをいくつ含んでいるか調査してみた。結果は表5の通りである。

縦のカテゴリーから検討してみよう。72%の特集記事が、その特集記事タイトルだけでは、それが予防に関係がある特集記事なのか、基礎研究なのか、診断なのか治療なのか、予防なのか、見当がつかないことになる。

横のカテゴリーでは55%がカテゴリー含有数2以下である。つまり、半分以上の文献が検索誤差を生じさせることになる。

これは、特集記事のタイトルが何件かの文献を代表した総合的なタイトルであることを考えると仕方のないことである。

3.4.3 抽出できないタイトルおよび用語

この抽出規準では次のようなタイトル、または用語が抽出できなかった。

イ) タイトル：改正診療報酬点数表

ロ) 用語：外来、職業

イ) は、縦のカテゴリーのいずれにも入れることができなかった。横のカテゴリーも同様であった。ここでは一応、診断、治療のカテゴリーを含み、それと操作が交差するものとして処理したが、本来から言えば、診断に

も治療にも関係のない概念である。

これは、外来とか職業とかいう用語を抽出できなかったことも含めて、このカテゴリー表が患者の生命を救うという医学の直接的な目的にのみ重点を置いて作られた結果であったと思われる。

基礎的な研究と診断、治療を同時にカバーするカテゴリー表を作るのにも、かなりの無理があったように思われるので、医療行為の社会性をも包摂するようなカテゴリー表を作成することは非常に困難なことであるが、次の機会の課題としたい。

4. 結 論

(1) このカテゴリー表で一応、医学の基礎、臨床部門の和雑誌特集記事のタイトルを主題分析出来るが、各カテゴリー間に抽出数のバラツキがあるので、次のカテゴリーを使用し、同種のデータで実験する必要がある。

イ) 横のカテゴリー

i : 疾病、傷害を併合する。

ii : 生体、物質を併合する。

iii : 生体構成部分を何らかの形で2つに分ける。

iv : 運動・作用・機能と状態・現象を併合し、機能・現象とする。

v : 性質・形態・程度と時間・空間を併合し、性質・形態・程度とする。

vi : 操作・方法・処置と操置・設備・材料を併合し、操作・設備・材料とする。

ロ) 縦のカテゴリー

看護を消去する外はそのままにする。

(2) 生体構成部分に抽出された主題の出現頻度をみても、概念レベルが高いことがわかるが、さらに研究・診療領域に属する主題が第3位にきており、具体的な研究・診療に使用する操置・設備・材料に属する主題が全体の1%しかないことを見ても、主題の概念レベルが高いことが判明した。総じて、各カテゴリーに抽出された主題は高い概念レベルを示した。

(3) 和雑誌特集記事のタイトルから抽出した主題だけでは、検索に非常に大きな誤差がでる。しかしながら和雑誌特殊記事タイトルの性格から仕方がないことであり、タイトル以外のところから出題を抽出する一方、利用者のニーズを調査し、そのニーズとのマッチングの実験をする必要がある。

(4) このカテゴリー表では、医療の社会的な概念を包含できない。それを包含するカテゴリー表を作成する必要がある。今後の課題としたい。

参考文献

- 1) 森 耕一：面に関する研究—物理学の場合「図書館学年報」4：9～31, 1957.
- 2) 斎藤 孝：索引作業のための自然語処理の研究 *Library Science*, 5：51～72, 1967.
- 3) Weil, B. H., et al. : Technical-abstracting fundamentals. I: Introduction, *J. Chem. Doc.*, 3(2) : 86～89, 1963.
- 4) Luhn, H. P.: Keyword-In-Context Index for technical literature. *Am. Doc.*, 11 : 288～295, 1960.
- 5) Kraft, D. H.: A Comparison of keyword in context indexing of titles with a subject heading classification. *Am. Doc.*, 15 : 48～52, 1964.
- 6) Montgomery, C.: Machine-like indexing by people. *Am. Doc.*, 13 : 359～366, 1962.
- 7) 斎藤 孝：索引作業のための自然語処理の研究。 *Library Science*, 5 : 51～72, 1967.
- 8) 坂本徹朗：主題分析 「情報管理」13 : 122～131, 1970.
- 9) *ibid* p. 128.
- 10) *ibid* p. 53.

(昭和50年8月13日受付)

国内ニュース

東 西 南 北

『雑誌記事索引』

—コンピュータによる機械編さん化—

国立国会図書館が、全国の学術雑誌の論文記事の索引誌として刊行している『雑誌記事索引』『人文・社会編』(昭和23年創刊)と「科学技術編」(昭和25年創刊)が1975年1月分から、従来の手作業による編さん方式からコンピュータによる機械編さん化に移行した。機械編さん化により、充実される面を挙げると

1. 現在刊行されている月刊版は、遡及的調査には不便なため、年間累積版の刊行が待望されていたが今後は容易に刊行し得るようになること。

2. 複合主題に関する論文記事は、関連する各主題項目に分類重出すると便利であるが、手作業方式では、索引カード記述の手数をかけられないため、著しく制限してきた。今後は機械編さんのため、同一データを必要な項目ごとに容易に重出できるようになること。

3. 著者索引、件名索引を両編ともに付する予定で、このような補助索引の充実も機械編さんの成果である。

しかし、コンピュータの増設、取替え等の種々の条件が時期的に錯そうするため、手作業による昭和49年12月号刊行後、50年分を引続いて刊行できない。事情が生じた。機械編さんによる『雑誌記事索引』の最初の号の刊行は両編とも、51年3月の予定である。

昭和50年分のデータのコンピュータへの入力はずでに開始されており、総合的編集プログラムの完成次第、蓄積データを累積版として刊行する。昭和50年分については、1—6月、7—12月の半年単位の累積版として続けて刊行する予定で、補助索引類は両編とも、1976年8月頃には別冊刊行される予定である。

昭和51年以降は月刊版と年間累積版をそれぞれ刊行し、期待に応えるよう準備している。

(「国立国会図書館月報」No. 172 : 19, 1975. 7)