Library Science No. 5 1967

索引作業のための自然語処理の研究
——医学用語の計量的調査——

A Study of Natural Language Processing for Indexing
—A Statistical Survey of Medical Terminology—

斉藤孝
Takashi Saito

Résumé

In relation with the problem of language processing for indexing and searching purposes, the writer carried out an experimental research of the processing of natural language in documents. The field of medicine was selected as the research subject, and the samples were taken from words appeared in titles of articles in the area of pathology.

In extracting terms from natural language, it is required to analyze the agglutinativeness of Chinese characters used in Japanese language. A Chinese character, as is well known, has its own meaning just like a word has. Consequently, the Chinese characters can be said to have word-construction function through which their different associations become to indicate different concepts. This causes a difficulty in identifying word division or unit word which is basic to the study of terms. Further, this is resulted in the ambiguity of structure expressed by the association forms of unit words.

To make clear the characteristics mentioned above, quantitative measurement by the statistical method may be of help. In this article, the writer tried to clarify, by this method, the nature of medical terminology and to evaluate its value as indexing terminology.

(Graduate Student, Japan Library School)

はじめに

ドキュメンテーションには，情報の蓄積と検索のための索引技術の開発がある。索引技術は，文献の内容の主題を分析することで，索引用語に変換する作業と考えられる。この作業は情報の蓄積と検索に当って，主題を一種のコトバによって処理加工することと言え，言語処理技術の場となる。そこでは，情報を特定の言語の記号列に変換すること，および，その言語記号列から情報を取り出すいわゆる索引作業（indexing）と探索作業（searching）とが必要とされる。

本稿では，この索引作業と探索作業という言語処理上の問題として，文献中の自然語の処理についての研究の結果を述べた。文献中の自然語と索引用語との関係は，次のような点にある。索引作業では，索引用語を自然語で書かれた論文や標題から抽出し，シソーラス等の索引用語集のコトバと比較することで，索引用語を選択，付加する。ここでいう自然語は，同義語，同形異語等の整理，さらに包括特殊関係などの概念処理などが一切加えられていないものを指す。そこで自然語の処理にはいわゆるシソーラス的な操作を必要とする。この種の操作に関しては，Costello[1]，Bernier[2]，Bourne[3]，Gillum[4]等が詳細に論じている。さらに実際的な点は，MEDLARS[5]等の索引マニュアルにおいて論じられている。現在，慶応義塾大学北里記念医学図書館では，アメリカ国立医学図書館（N.L.M.）での MEDLARS 計画の一環として日本の医学文献に対しての索引作業を行っている。[6] この作業に関し，日本語による医学文献に対して，英語の索引用語集である MeSH から索引用語を付加する上で，次の問題点の解決が必要となる。それは，日本語という自然語により記述された医学論文は，いかなる傾向の用語を使用しているのか，そして，それらから索引用語を抽出する場合の問題，さらに MeSH のような英語の索引用語との関係はどのようなものであるか等の問題である。この点に中心を置いて，医学用語の実態の調査が試みられた。[7] 本稿では計量的な側面から，この調査法についてさらに検討を試みた。

I. 調査対象の設定

A) 標本の採集

医学における諸主題は非常に広い領域を占めるものであり，したがって本調査では「病理学」に対象を限定した。病理学の領域は，次のごとく簡単に定義することとする。「生物学に属する一学科で疾病及び病的状態に関する事項を研究するものであり，その中でも疾病に際して現われる形態学的変化に就いて研究するものを病理解剖学と名付け，特に組織的変化を研究するものを病理組織学と言う。その他，病理生理学や化学的変化を追求する病理化学などがある。」[8]

病理学領域における用語は，論文標題から抽出することとした。標本論文は，抄録誌である医学中央雑誌[9]の分類件名〝病理解剖学〟の下から採集した。医学中央雑誌は1ヶ月に5冊発行され，6冊をもって1巻を形成するもので，1ヶ年で10巻が刊行されている。調査は，第172～182巻（昭和37年度）の1年間に収録されたものに限定して行った。なお，昭和37年度（1962）において収録された総論文抄録数は1,837であり，ここから任意に採集した論文標題1,000を標本とした。したがって，母集団は病理学領域での全標題をいちおう代表するものと考えられる。

B) 標題用語の価値

調査対象とした用語は，論文標題中に現われたものに限定されている。したがって，論文標題をどのように評価するかによって用語の価値が変ってくる。元来，〝標題〟とは論文の内容を代表するものと見なされる。この点に注目した索引方式として，KWIC[10]が有名である。KWIC 索引方式におけるキーワードとしての標題用語の価値は，次のような理論的根拠に支えられていると言えよう。すなわち，コトバは文脈というある与えられた状況の中で，常にただ1つの意味を持つものであり，それにただ1つの概念が対応するものである。そこに文脈的意味という捕え方が可能となるが，KWIC ではそれを標題のレベルに依存していることになる。したがって，論文標題が最も論文主題を忠実に反映したものであるという前提に立つ仮説によっていると言えよう。この仮説に対しての実証例として，英文文献では法律領域での Kraft[11]と医学領域での Montgomery[12]等の調査がある。しかし日本語文献での例は，まだきわめて少数に止まる。そこで今回の調査でも，目的の一つとして日本語の標題用語の妥当性の評価をも試みた。なお，医学論文標題の例には，次のような形式がある。

①実験的 Toxoplasma 感染症の脾，脾静脈，門脈に於ける白血球の消長に関する研究。

②Virus 性腫瘍の本態に関する実験的研究（2） 誘発に於ける家鶏肉腫 Virus の形成過程。

③障碍細胞の病理形態学的研究（2） 実験的に結紮され

た頸動脈の電子顕微鏡的観察。

④組織培養に由るラッテ腹水肝癌細胞の研究（10）浮遊状培養に由る組織培養株細胞の染色体の変化。

これらの主題用語は実際に使用される，いわゆる〝なま〟の用語であることに特徴がある。標題用語の価値は，その妥当性は別として，索引作業においては索引者が主題分析上の目安とすることは明らかな事実であるから，ともかく標題用語の実態を調査しなければならなくなる。調査では論文標題の価値を具体的に，次のような仮説のもとに評価した。

①論文標題は本文の長さの 1/100～1/1,000 である。

②さらに抄録の長さの 1/10 である。

③標題用語は多くの索引用語（キーワード）を含む。

④標題は「抄録の抄録」と言われているように，情報密度の点で優れている。

⑤したがって標題用語はKWIC索引方式に利用できる。

⑥標題の用語が構成する形式は，一般の文章よりも単純であり，日本語から英文標題への翻訳が比較的に機械的に行なえるものである。

これらの仮説は，実証されなければならないが，このことについては，後にキーワードの評価として論じた。

II. 用語の抽出方法

文章の構造に関し，主語，述語，目的語のように順に単語を並べることで，それらの相対位置によりその機能が自から決まる欧文の場合とことなり，日本文の場合は単語の区分が不明確である。単語と単語との間に，英語のようにスペースが設けられていない。特に日本語の文章では，助詞，助動詞という独特のものがあって，語の区分形式をあいまいにし，さらに接頭，接尾語が多いが，これらの語は用語をつくるのに重要な働きを示す場合が少なくない。また，漢字による造語の性質は，造語過程の分析をしなければ明らかにならない。したがって用語抽出上に当っても，その単位となる基準を明確にする必要がある。普通，コトバ(語句)の単位と考えられるものに分節というものがある。しかし漢字を使用する場合では，重ねて結合していくことで非常に長い一つのコトバを作ることが可能となる。いわゆる漢字の使用に見られる膠着性により，どこまでを一つの単語又は単位とするかが明確でない。単位としては，極端な場合では漢字一字とすることも考えられる。したがって，単位というのは非常に伸縮自在なものであることになる。調査では便宜上単位を決める上で，次に示す認定単位を利用した。

A) 認定単位の設定

用語の抽出基準となる認定単位は，次のように規定した。

①非重要語：助詞，助動詞，接頭語，接尾語等。

②合成語：次に述べる，単位語と単位語，又は単位語と特殊単位語との集合によるもので一つの特定の概念を示す用語（term）又は句（phrase）を形成する語。例えば〝甲状腺腫，十二指腸虫，病理学的研究〟。

③単位語：ⓐ漢字2字の集合で，一つの特定の概念を示すことの出来る語。

ⓑ単位語中の常用語とは，特定主題領域でつねに生起する語および，それ自体，意味的には重要な価値を持たず，論文の形式，または慣習的に付加する語で，例として〝所見，剖検，研究症例，検討，展望〟などがある。常用語は，後に論ずるキーワードの選択に関係する。

④特殊単位語：ⓐ現代医学用語中，漢字一字でも用語（term）として認められるもの。例えば〝胃，肝〟

ⓑ補助語となるもので漢字一字で接尾的な機能を果すもの。例えば〝～的，～様，～型，～状，～体〟などと〝～病，～学〟など。

ⓒ外国語で，英独語の用語が大部分である。また例外的に漢字2字以上のもの，例えば〝顕微鏡〟などのごとく，分離することにより本来の概念を全く示すことの出来ない語。

以上のものを認定単位とする。これらの単位を抽出する過程，すなわち認定操作は第1図に示したフローチャートによる。このフローチャートでは，単位語のリストを最終的に求めるものとなる。なお除外される語としては非重要語と数値，固有名，特殊記号，等である。

B) 認定操作の方法

1) 合成語の抽出

標題1,000から標題を構成する語のうち，助詞，助動詞，接頭，接尾語等の非重要語を除外したものを2字以上の語の結合した状態のまま抽出する。次の標題：

①〝細胞分裂の調節機構に関する研究〟

②腫瘍移植性に対する純系動物の病理学的検討〟

③〝Ehrlich 腹水癌の マウス 皮下移植結節から抽出した Oncotrepin と Antitrephin〟

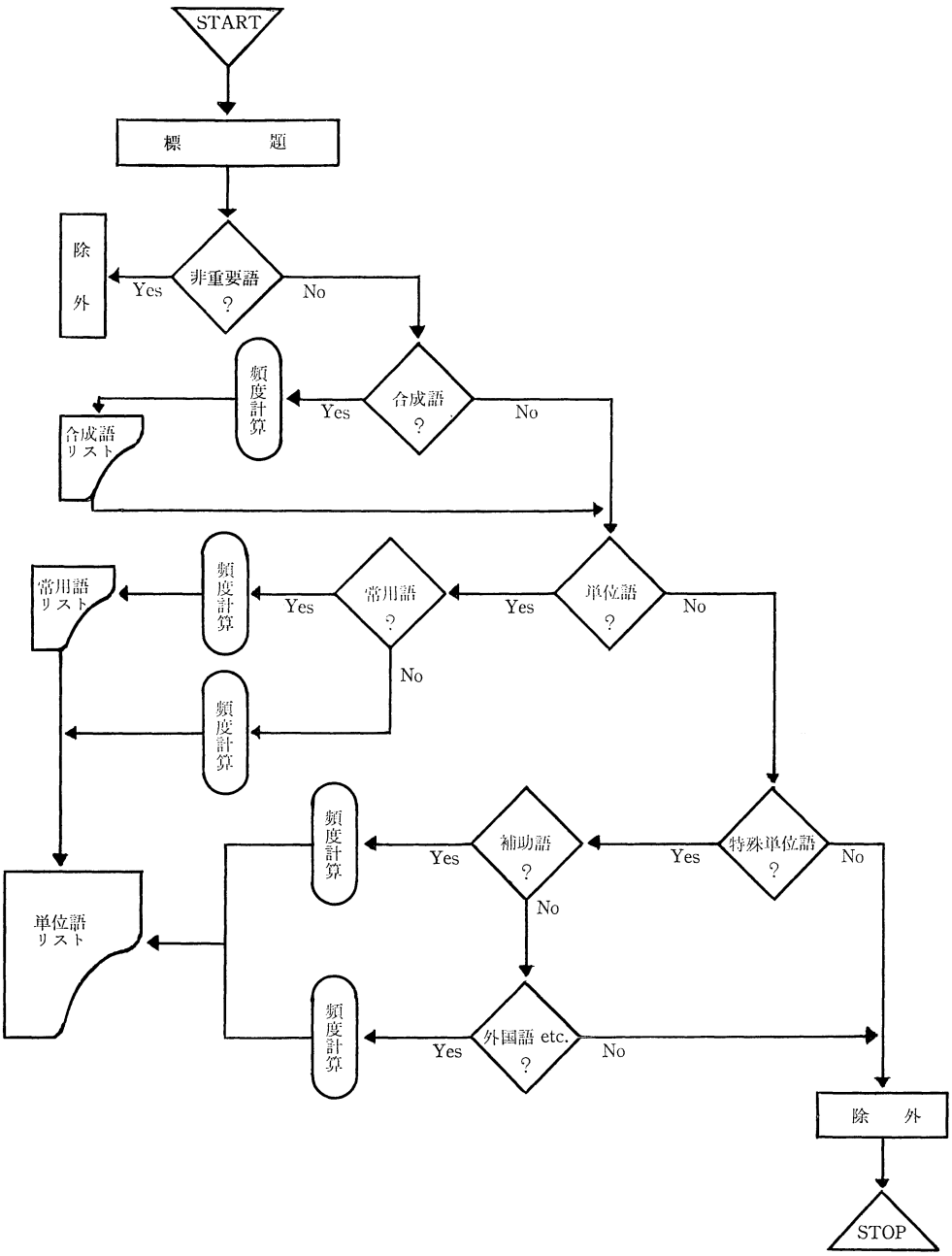
等では下線の語が除外されて，残ったものが，そのまま合成語となる。合成語の特徴は，日本語に特有なものである漢字の造語力によって出来ている点に見られる。漢字に元来は一字でも，そのまま一語としての意味を持つ

ものであり，したがって漢字が集合することで異なった概念を示すコトバを形成することは比較的簡単なものとなってしまう。これらの合成語は，頻度計算により合成語リストを作成して，後に論ずる結合力という尺度で分析する。

2) 単位語と特殊単位語の抽出

単位語は便宜上の名称にすぎない。何故ならば先にも論じたように，日本語では漢字の造語力が強いことによ

第１図　認定操作のフローチャート



り大部分は合成語の状態で表記されるからである。したがって，どこまでを単位語とするかを厳密に規定する手段がとれないことになる。英語のように単語を単位とした〝分かち書き〟をしていないので，一つの単語を単位区分として設定することができない。この調査でも，単位語の認定は外務省の方針[13]に見られる方式を利用している。それは次のようなものである。

ⓐ文を単語で切る（文節から付属語を切り離す）。これは合成語の抽出方法と同じもの。
ⓑ接頭語，接尾語を切り離す。
ⓒ数をあらわすものは一つの単位とみなす。
ⓓ3字以上の漢字は，慣用的に見てその部分がまとまった概念を表わしていれば，2つ以上に分ける。

そこで，合成語からの単位語の抽出は，〝細胞分裂〟で

リスト1　単位語のリスト

順位	単位語	頻度	%	順位	単位語	頻度	%
1	*～的	368	4.41	33	結核	41	0.49
2	癌	286	3.43	34	*～剤	41	0.49
3	*～学	248	2.98	35	移植	38	0.46
4	*～性	245	2.94	36	*～例	38	0.46
5	研究	241	2.89	37	甲状	37	0.44
6	組織	232	2.78	38	ラッテ	37	0.44
7	細胞	210	2.52	39	*～酸	36	0.43
8	病理	182	2.18	40	転移	34	0.41
9	腫瘍	144	1.73	41	反応	33	0.40
10	肝（臓）	134	1.61	42	障碍	32	0.38
11	*～症	125	1.50	43	白血	32	0.38
12	肺（臓）	92	1.10	44	染色	32	0.37
13	化学	71	0.85	45	臓器	30	0.36
14	マウス	69	0.83	46	Virus	30	0.36
15	顕微鏡	65	0.78	47	淋巴	29	0.35
16	脳	65	0.78	48	硬変	27	0.32
17	電子	65	0.78	49	動物	26	0.31
18	*～腺	60	0.72	50	観察	26	0.31
19	実験	59	0.71	51	増殖	25	0.30
20	肉腫	58	0.70	52	酵素	24	0.29
21	*～炎	57	0.68	53	*～像	24	0.29
22	変化	55	0.66	54	病変	23	0.28
23	*～系	51	0.61	55	*～法	23	0.28
24	形態	50	0.60	56	中毒	23	0.28
25	腎（臓）	49	0.59	57	硬化	23	0.28
26	動脈	48	0.58	58	物質	22	0.26
27	発生	47	0.56	59	*～類	22	0.26
28	*～体	46	0.55	60	神経	22	0.26
29	胃	43	0.52	61	所見	22	0.26
30	培養	42	0.50	62	家兎	22	0.26
31	腹水	41	0.49	63	多糖	21	0.25
32	血管	41	0.49	64	異常	21	0.25

＊　特殊単位語（補助語）
頻度順位 64 位まで

は〝細胞〟と〝分裂〟が，〝結核病〟〝脳腫瘍〟では〝結核〟と〝腫瘍〟が分離される。また後者においては〝～病〟と〝脳〟は特殊単位語と認定されることになる。

3) 頻度計算

頻度計算は同一語としての認定操作による。同一語という概念は，表記体としての〝意味〟も〝形式〟も同じものを指す。したがってこの認定は，文脈中での単位語または合成語中での単位語によるものではなく，抽出された語と語との照合による。たとえば単位語の集合である〝発癌剤〟と〝腸癌〟とでは，単位語〝癌〟のレベルでは同一語と認定されるが，合成語としては全然別の語となる。ここでの問題点は，例として〝Allergy 性〟と〝性科学〟に見られる〝性〟の場合などの同音異義語又は同形異義語の処理に関して現われる。この点の分析はsemantics, syntax によるアプローチを必要とし，純粋に表記体での〝形式〟だけからは決め難い問題となる。この事実は，〝意味〟を無視した結果が，はたして索引語を求めるための用語調査として受取られるかどうかという問題へと発展する。

そこで，この処理に関しては一つの作業仮説のようなものを考えた。それは，〝意味〟は〝形式〟よりも客観性に乏しく，内省にたよらなければ捕えにくいし，それ故に極端な場合では，たとえ表記体の〝形式〟では同一語となっても，人によっては全然相違した〝意味〟を示していることも考えられる。さらに，単位語のレベルでは同形異義語を判定出来るものではない。この調査の特色としては，単位語を合成語，または文脈中で捕えてはいない点に見られる。したがって，〝意味〟による適合の概念は全く適用されない。以上の認定操作により次の概念を引き出した。

①母集団：調査の対象となった，病理解剖領域における論文標題全てを言う。

②標本：調査で対象とした母集団より任意に採用した論文標題1,000を言う。この標本は用語の集団と見なされる。したがって単位語，特殊単位語，合成語の集合 $<U>$ と考えられる。そこで単位語と特殊単位語は，集合の要素 $<W_n>$ となる。集合 $<U>$ は要素 $<W_n>$ としての単位語が表記体上の〝形式〟のみにおいて同じ姿を示すもの，すなわち同一語の集まりを意味する。その関係は "$U=(W_1, W_2, \ldots W_n)$" となる。

③生起頻度：母集団または標本中において認定される同一の単位語，すなわち $<W_n>$ の使用される回数を意味する。

④延べ語数：標本中において認定された，すべての単位語の総和。

⑤異なり語数：標本中において異るものと認定された単位語の総数。用語の語彙とは，標本中でそれぞれが異ると認定されるすべての単位語の集合となる。

以上の用語の規定により次のデータと単位語のリスト1が作成される。

標本標題数	1,000
単位語延べ語数	8,037
平均標題単位語数	8

III. 用語の分析と評価

A) 分析方法

分析は単位語，合成語，特殊単位語に分けて行う。

1) 単位語の結合力の分析

単位語の分析は，標本中での語彙量の測定に関する項で詳細に論じる。ここでは，単位語と単位語との関係の分析手段として結合力という尺度について論じたい。単位語は用語の大部分を占める合成語の構成要素となるものであり，したがってこの要素を基礎として量的関係を捕えることが試みられる。つまり基礎となる単位語の前後に集合する単位語，または特殊単位語の集合数によって，いかなる単位語が，いかなる割合で合成語を形成するのかを表示しうる。そこで結合力は，リスト2に示した例に見られる基本の単位語を中心にした関係により，次のように表示される。

$$c_{fw} \fallingdotseq \frac{w_r}{f_b}$$

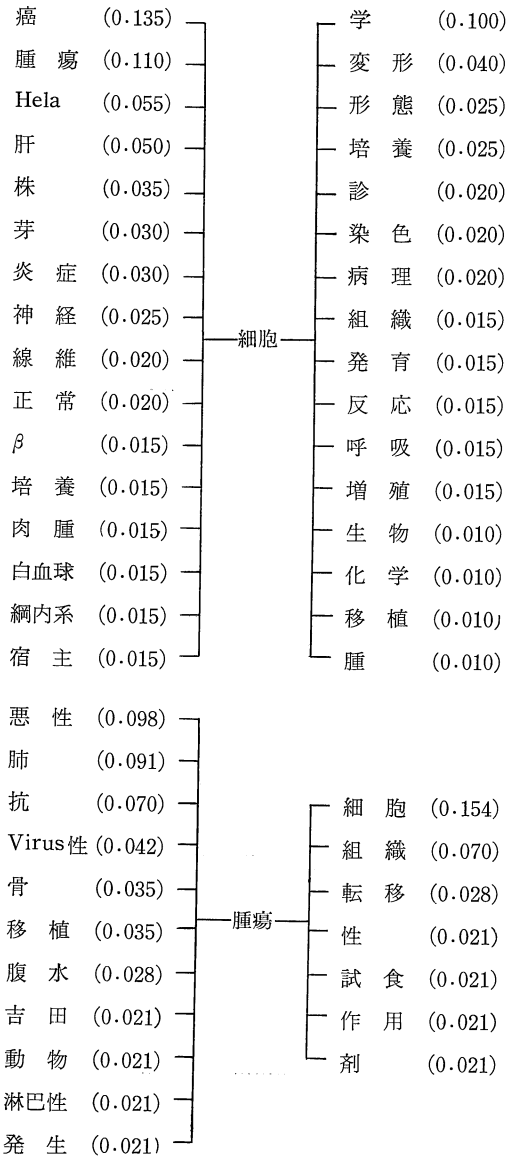
c_{fw}＝結合力　f_b＝結合単位語の集合回数

w_r＝基本の単位語の標本中での生起頻度

2) 合成語の構造形式の分析

合成語は単位語または特殊単位語の2つ以上の集合したもので，語 (term) と句 (phrase) を形成するものである。単位語と比較すると，合成語は〝癌＞腹水癌＞ラッテ腹水癌〟など〝癌〟が結合した場合概念の大小を示すように，語の意味の領域を限定する働きをする。合成語は，単位語の集合による結合構造による重要な機能を持つ。特に漢語の造語力又は造語過程と言われるものによる単位語の膠着性の分析を研究する上で，合成語は材料となる。結合構造に見られる単位語の連体修飾の関係を，ここでは考慮してみよう。たとえば〝腫瘍細胞形態

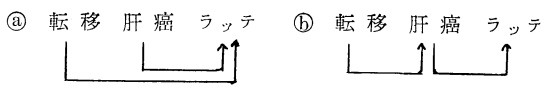
リスト 2　結合力のリスト



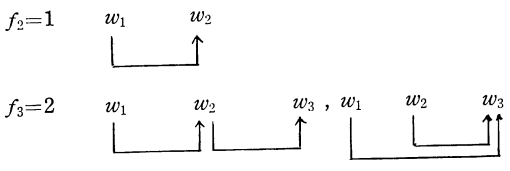
学的検索〃，〝完全内蔵反転症〃〝全身性類粉変性症〃等の合成語は一つの用語として判断するものでなく句と考えるべきものとなるが，この種の句を分析する時，意味を度外視しても，単位語と単位語との結合関係による，構造上の意味の特殊性が見られる。この種の語句の造語過程には，先に論じたように漢語の膠着性が見られ，極端な場合では，その合成作用が無限に近い形で延長され，それにより特定の〝意味〃を表示できるという独特な構造を有する。この構造的に生ずる意味を，結合構造に見られる単位語の結合する関係を捕えて計量的に分析を試みようとする方法がある。[14] この方法は，文法における助詞〝の〃を用いる連体修飾の関係を，その構造形式により説明しようとするものである。合成語と見られる句については，同じことが考えられるものとして利用してみた。今，単位語を“w”とすると，合成語は“$w_1+w_2+w_3+w_{n-1}+\cdots+w_n$”となる。ここで“$n$”は，単位語の膠着性により，任意の数と考える。例として合成語〝転移肝癌ラッテ〃は単位語では〝転移〃，〝肝〃，〝癌〃，〝ラッテ〃の集合となる。仮に結合構造が明らかでない場合は，次の解釈に基く。

ⓐ〝転移〃した〝肝癌〃の〝ラッテ〃⟶転移肝癌ラッテ
ⓑ〝肝癌〃を〝ラッテ〃に〝転移〃⟶肝癌ラッテ転移
ⓒ〝肝癌〃を〝転移〃した〝ラッテ〃⟶肝癌転移ラッテ
ⓓ〝ラッテ〃に〝転移〃した〝肝癌〃⟶ラッテ転移肝癌

ここでは，特殊単位語〝肝〃と〝癌〃とは，合成語〝肝癌〃とみなす。上記の例では，合成語を単なる単位語の羅列と解釈したことに基づく結合関係のあいまいさが生じると考えられる。しかしこの点では，日本語の文法では問題を生じない。なぜならば後に出て来る名詞が，先行する名詞を修飾しないという規則があるからである。それ故に，考えられる結合構造でのあいまいさは次の2つとなる。



この例で矢印は，単位語の結合関係を示す。この場合，いわゆる語と概念の連合による差違に注意しなければならない。たとえば，概念として考えた場合，ラッテは転移しないが肝癌は転移するのであってⓐのケースは考えられないものと言える。そこで構造の型，すなわち構造形式を分析するための手段を考えてみる。次に示す単位語を w として，その構造形式の数 f_n は，合成語“w_1, w_2, w_3,$\cdots w_n$”では，

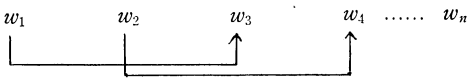


f_n を構造形式の総数とすると矢印のつけかたにより，この合成語の構造の種類が決る。ここで日本の文法規則に

より次のような限定を作る。

①矢印は必ず左から右に向けてつける。つまり結合関係での方向が w_i から w_j へつけてあれば“$1\leq i<j\leq n$”である。

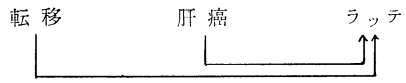
②2つ以上の矢印が互に交叉することは許されない。つまり $w_i\rightarrow w_j$ と $w_j\rightarrow w_i$ と言う3つの結合関係での方向は“$1\leq i<i\leq i'<j'\leq n$”かまたは“$1\leq i'<j'\leq i<j\leq n$”である。すなわち



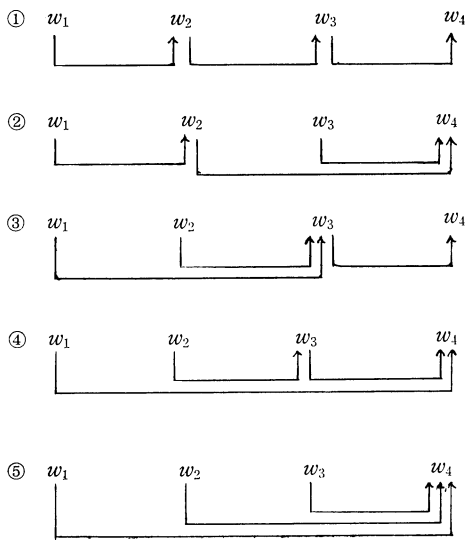
の構造は許されない。

③“$w_1, w_2, \cdots, w_{n-1}$”の結合においては，一方向の矢印だけが出る。この限定は，1つの w はただ1つの w だけを直接修飾して2つの w に同時に直接かかってはいけない，すなわち矢印が1つの単位語から2本出ることはありえない，ということを意味する。

④“$w_2\sim w_n$”では1つの w に同時に何本もの矢印が入って来てもよい。この限定では1つの w がいくつかの w によって直接修飾されていてもかまわない事を意味する。



以上の限定では，例として“$n=4$”では，合成語の構造形式の数は“$f_4=5$”となる。

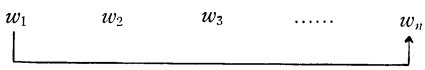


したがって今 n 個の w すなわち“$w_1, w_2, \cdots, w_n$”に結合関係をつけるとすると，

$(n-1)$ 個の w につけた矢印の数 f_{n-1} となる。

②$w_1, w_2\cdots w_{i+1}\cdots w_n$

w^1 から出た矢印が w_{i+1} $(1<i+1<n)$ を指している場合では，$w_2\cdots w_i$ の矢印は w_{i+1} を越すことができない。これでは構造関係の数は f_{n-i} となり構造関係の総数は $f_i\cdot f_{n-i}$ となる。



w_1 から出た矢印が w_n を指している場合では，残りの矢印は w_2 から w_n まで $(n-1)$ 個の w につけた矢印と考えられる。その構造関係の数は f_{n-1} となる。

したがって①，②，③，を総合すると

$$f_n=f_{n-1}+f_2\cdot f_{n-2}+f_3\cdot f_{n-3}+\cdots+f_{n-2}\cdot f_2+f_{n-1}$$

となり $f_1=1$ と定義すると

$$f_n=\sum_{i=1}^{n-1} f_i\cdot f_{n-i}\ (n\geq 2)\cdots\cdots(1)$$

ところで巾級数

$$F(x)=f_1x+f_2x^2+f_3x^3+\cdots\cdots$$

において (1) 式を使えば

$$\{F(x)\}^2=F(x)-x\cdots\cdots(2)$$

この (2) 式を $F(x)$ について解けば，根の一つとして

$$F(x)=\frac{1}{2}(1-\sqrt{1-4x})$$

が求められる。これを巾級数に展開したときの係数が (1) に等しく，その係数 f_n は

$$f_n=\frac{(2n-2)!}{n!(n-1)!}\cdots\cdots(3)$$

となり，(1) よりも簡単に f_n が計算出来る。例えば $f_1=1$, $f_2=1$, $f_3=2$, $f_4=5$ となる。

次に (3) を変形して

$$f_{n+1}=\frac{2(2n-1)}{n+1}\cdot f_n \quad \text{とすれば}$$

$$\lim_{n\to\infty} f_{n+1}=\lim_{n\to\infty}\frac{4-\frac{2}{n}}{1+\frac{1}{n}}\cdot f_n=4f_n \quad \text{となる。}$$

従って n 個の単位語 w の結合による合成語の構造形式の種類 f_n は，単位語の個数 n が1つ増える毎に大体

4倍に近づくと考えられる。この構造形式の分析は合成語の〝意味〟に関するsyntax上の分析に対する一つの手段と言えよう。

3) 特殊単位語の分析

特殊単位語の特色は，外国語を直接用いる場合と例外と見なされる語以外は，一字の単位で特定の概念が認定されることである。これらは，現代医学用語として漢字一字でも用語(term)として認められているもので，例としては〝癌，骨，腸，腎，肝，脳〟，等が挙げられる。一方，補助語といわれる認定単位は，同じく漢字一字であるが，接尾的機能を果す語として〝～性，～化，～的，～状，～学，～系，～様〟等が挙げられる。この分析では，この種の補助語のみに限定して評価を試みた。補助語の特徴は，それ一語では具体的な概念を示すことが出来ないことにある。それらは，合成語中で機能的な働きを示す語と考えられる。〝可移植性〟，〝淋巴性〟，〝紅斑性〟，〝抗原性〟，〝Allergy性〟，等の〝～性〟は性質を示す機能語となる。医学用語では補助語の多用が大きな特色となっていることは，抽出数が多いことでも明らかである。そこで，補助語は合成語中では，いかなる機能を果すかを分析してみたい。例としては〝～状，～的，～様，～性，～化〟等の頻度の高いものを挙げてみた。

①〝～的〟：最も生起頻度の高い単位語であるこの〝～的〟の意味の働きが不明瞭である事が多い。つまり機能領域をはっきり示すことがむづかしい。〝～的〟は「らしさ」という性質，ある傾向にあることを示し，その意味の機能は「らしさ」，「疑似性」を示すものである。一方，中国語の「的」に等しく助詞「の」の働きをする格助詞となって，連体修飾語を作る機能がある。したがって，慣用に従いいかなる単位語とも結びつきやすくなる。特に膠着性造語過程では，形容詞化の機能を果す。

〝実験的肉芽組織形成〟，〝病因的考察〟

〝電子顕微鏡的細胞組織〟等がその例である。

また，抽象度の高い他の単位語と結合する傾向が見られる。

②〝～性〟：性質または属性を表現する機能を果す。〝Allergy性肺炎〟，〝進行性背髄性筋萎縮〟，〝Virus性白血病〟は〝～的〟とほぼ似た傾向をもち，大部分の単位語に結合する。しかし，次のような場合では，意味的差違が明らかではない。

① 淋巴肉腫 / 淋巴性肉腫　② 異種腫瘍細胞 / 異種性腫瘍細胞

④〝～様〟，と〝～状〟：いずれも様子，状態を示す機能を果すもので，形態の描写に多く使われる。〝絲状，花房状，葡萄状，星状〟，あるいは〝Virus様，腫瘍様，皮腫様，リウマチ様〟等に見られる傾向としては〝～状〟では物質名を示す単位語との結合力が高く，〝～様〟では病名，現象を示す単位語との結合力が高い。

④〝～化〟：作用や変化を示す機能を果す。〝四塩化，石灰化，石炭化，繊維化，安定化，実験化，活性化，異型化，癌化，腫瘍化〟等に見られるように〝～性〟，〝～的〟と同様に比較的，いかなる単位語とも結合しやすい。

これらの〝～的，～性，～化，～様，～状〟等の補助語に共通する傾向としては，臓器名を示す単位語，たとえば〝心臓，胃，肝，腎，脳〟等と結合する例が非常に少ないことが挙げられる。

B) 医学用語の評価

単位語リストと，そのデータに基き，標本として採用した論文標題用語の評価を試みた。標本である医学中央雑誌（昭和37年）1年間に収録されたものから，平均的な標題として「癌細胞組織の病理学的研究」を推定した。これは，標本中での生起頻度上位の単位語とその結合力により考えられたものである。医学用語の特質としては，用語の使用法が非常に粗雑で難解なことが挙げられる。たとえば単位語の無制限に近い集合による造語と，さらに補助語の機能を乱用した〝x化β状α性μ的〟等の構造形式などである。次のような例では，語なのか句なのか，さらにどれを単位語とすべきかが判断しかねるものが非常に多く抽出された。

〝悪性型黒色棘細胞増殖症〟

〝直腸周囲組織異所寄生〟

〝電子顕微鏡的細胞組織病理学的研究〟

この種の合成語は，日本語の表現法とは遠く，むしろ中国語的なものと言える。この原因の一つとしては，近代医学が西洋から技術や考え方を多く取り入れたことにより，外国語による多くの概念を示す場合に，無理に漢語による表現を取らざるを得なかったことが考えられる。すなわち，ここに西洋の概念と中国の漢字との奇妙な結合を産むことになり，多くの不思議な造語を生む結果になったと想像される。他方，医者，学者などが自分の説を論ずる場合に，ことさら見識を示そうとして普通の用語を使うのを嫌って，独特の用語（jargon）の乱用を考えだす傾向も見逃せない。そこで，さらに奇妙な漢字の

結合による合成語が，競って使用されだしたとも考えられる。この無批判な漢語の使用による造語で，奇妙な学術用語が誕生することになり，同じ主題領域でありながら，相互に意味の解せない用語が頻繁に使用される結果となる。

IV. 語彙量の測定

単位語の性格の分析は，単位語の示す外延，内包的意味を基準とするカテゴリー化によるアプローチと，標本中の生起頻度による Zipf 法則，確率分布，等を利用する統計的アプローチによるものが試みられる。この2つのアプローチは，単位語の集合としての標本の語彙を評価すること，すなわち，語彙量を測定する尺度として利用される。語彙量の測定は，特定主題領域での用語群の一つの性格を表現するもので，索引用語(キーワード)の選択,抄録の作成での重要な手がかりとなるものである。ここでいう語彙量は，単位語としての異なり語の，母集団や特定標本中の生起パターン分布で測定できるものとする。

A) カテゴリー化による尺度

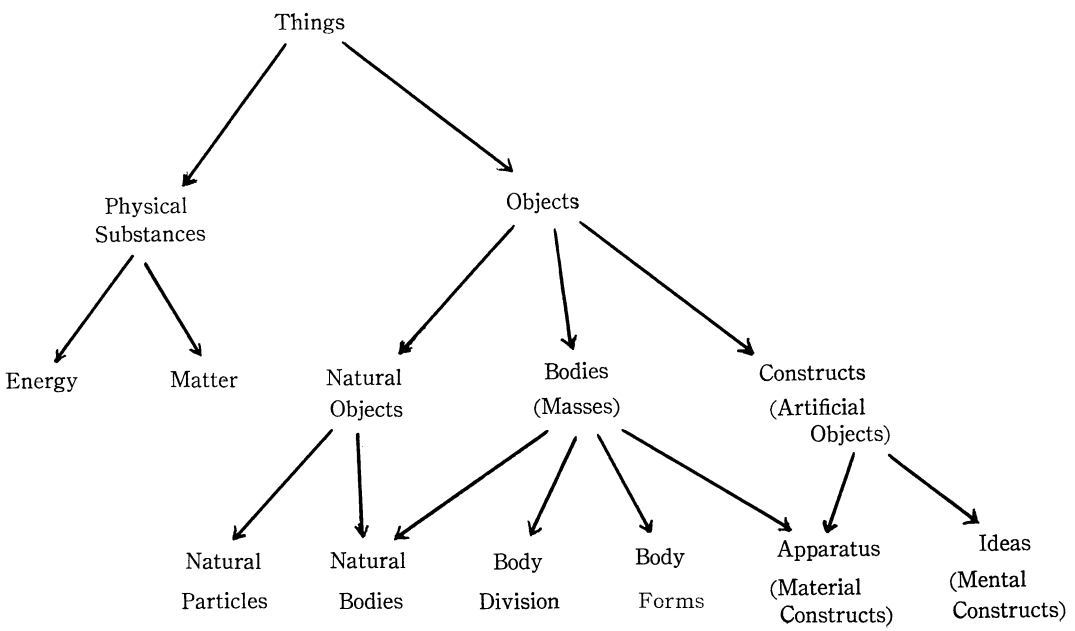
単位語のカテゴリー化は，前提として個々の語の意味の分析と概念の関連づけを必要とし，困難が多い。そこで総体的な主題領域の把握により，次のような量的な基準を考える。標本中での単位語の生起頻度に基き，上位の単位語の群を観察し，次のような平均的標題「癌細胞組織の病理学的研究」を推定した。さらに各単位語の結合力により〝胃癌，腹水癌，肝癌，肺癌，膵癌，乳癌，発癌物質〟等の中心主題を捕える合成語で主要概念を分析する。次に単位語の意味規定による類を設定し，いわゆるカテゴリー化を行う。類の設定により，単位語の性格を知る尺度が求められる。それにより，数量では表示できない単位語の意味的な分散状態を知ることができる。

カテゴリー化には演繹的方法と帰納的方法との2つの立場が考えられる。

①演繹的方法：一定の領域で使用される単位語の母集団またはその特定標本中で，単位語は，自然発生的に生れでたものであり，したがって，単位語の示す概念は，必ずしも特定の知識とか学問領域には平均的に出てくるものではない。そこで，体系分類に見られるような，ある特定の知識領域を限定し，単位語の示す概念をあらかじめ設定されたカテゴリーにくりこんでいく。例としては次のようなカテゴリーを利用する。

A) 対象（動物，ヒト）

第 2 図



B） 方法論（相関的研究，分析的研究）

C） 環境条件（in situ, in vitro, 低体温，低酸素，異常）

D） 研究目的方法（血行動態，代謝，調節機構，血液性状形態）

E） 対象臓器（心臓，腎，脳）

②帰納的方法：特定の領域の単位語の群を任意に捕え，その中に同一のカテゴリー化に可能な要素を求めていく。この方法で求めようとするカテゴリー化は，Vickery[15]による〝性質，材料，用途，プロセス〟，等のfacetの設定での手法が参考となる。彼は〝もの(thing)〟の属性に注目してグルーピングを行っている。この種の例では，第2図のような展開を考える。[16] 帰納的方法はグルーピングの手法とも言える。単位語のカテゴリー化も，この方法を利用して行った。標本中の異なり語より任意の個数の単位語を拾い出し，その意味に発見される一つ，あるいはそれ以上の要素がすべてに共通している語だけを取りあげてみる。次に，共通要素を少なくともその属性の一つとして持つ類を選び，その類に属するあらゆるメンバーを〝もの〟として規定できる概念で規定していく例としては Ⓐ器官(心臓) Ⓑ問題(結核) Ⓒ症状(発癌) Ⓓ能因（ヴィールス） Ⓔ処理（手術）等となる。

このカテゴリー化は，語の定義の解釈が中心になるので言語学的手法が強すぎ，従って客観性に不足するのが欠点となる。以上の方法を参考にして，次のカテゴリーを設置してみた。Ⓐ(器官，器具)，Ⓑ(症状，現象)，Ⓒ(用途，性質)，Ⓓ(対象，能因)，Ⓔ(操作，処理)である。このカテゴリーに属する単位語は第1表のようなものとなる。

第 1 表　カテゴリー別単位語

A(器官，器具)	B(症状，現象)	D(能因，対象)	E(操作，処置)
肝	癌	マウス	培　養
肺	腫　瘍	ラッテ	移　植
脳	肉　腫	家　兎	転　移
動　脈	結　核	蛔　虫	反　応
胃	中　毒	Virus	投　与

* C（用途，性質）は補助語，〜性，〜的の付加する合成語となる。

カテゴリーに属する単位語の数によりカテゴリーの大小を示すとⒶ22%，Ⓑ16%，Ⓒ25%，Ⓓ13%，Ⓔ23%であり，したがって“C>E>A>B>D”となった。これが単位語の類での分散を測定する尺度となる。

B) 統計的アプローチによる尺度

統計的アプローチによる語彙量の尺度は，本稿の中心となるものである。統計的という意味は数量的とか，多くのものの集団とかに関して，言わば確率的という概念で解釈する事に等しい。したがって単位語の語彙の性質を知る尺度としての数値が必要となる。数量化への第一歩は，どこに基準となる数量的計測単位を見つけだすかである。これは，事象の主要な条件や因子をいかに数値に対応させるかによる。さらに発見された数値的関係を尺度に変換して，測定という手段を見出すことで意味を持つ。ここで考えた測定とは，事象の数量的記述であると言える。測定との関連において問題となる数値は，〝同一性，順位性，加算性〟である。母集団または標本中での単位語の量と状態との把握は，これらの数値の性質を尺度としたものとなる。本調査における語彙量の測定は，この尺度を公式化することによった。

1) 単位語の量と生起分布

調査の結果，標本中での単位語の生起頻度とその使用率を示すリスト1が得られた。リスト1は単位語の上位64語を示した例であるが，これらの語は標本1,000論文標題に対して抽出された延べ語数8,037語を認定操作を経て異なり語1,104語としたものである。その関係は第2表である。

第2表の関係を単位語としての異なり語（x）と，その使用率の占める割合（y）で捕えると，第3表のように示される。さらに第3表の関係を x, y 軸の座標にプロットすると第3図のようになる。

第3図の各点を両対数グラフに目盛ってみると，ほぼ直線になるので，第3図の曲線を表わす函数として次式を考える。

$$y=ax^b \quad x>0 \quad 1{,}104\geq x>0 \cdots (4)$$

(4) 式における未知数 a, b は，

$X^d=\log x$, $Y^d=\log y$, $A^d=\log a$

として $Y=A+bX$ から求められる。よって第3図の曲線を表わす式は近似的には，$y=19.55\,x^{0.236}$ $1{,}104\geq x>0$ となる。この式は単位語の標本中の性格，すなわち x, y 間での相関パターンを示す。さらに母集団での単位語の量を表現する。次に単位語の生起頻度とその順位とを，y, x として両対数グラフにプロットしてみる。これにより単位語の生起

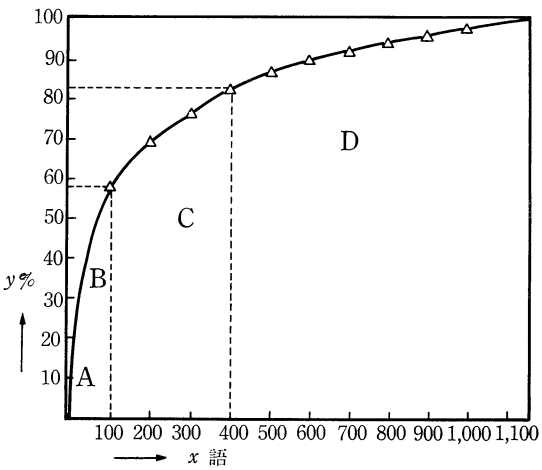
第 2 表

	単位語	*特殊単位語	総 数
延べ語数	6,188	1,847	8,034
異なり語数	854	250	1,104
**平均使用率	7.3	7.4	7.3

* 特殊単位語 {補助語 37 / 外国語 213}

** 各異なり語の平均使用率

第 3 図 単 位 語 の 量



第 3 表

*x (語)……y (%)
100 …… 58
200 …… 70
300 …… 76
400 …… 83
500 …… 86
600 …… 89
700 …… 92
800 …… 95
900 …… 97
1,000 …… 98

**単位語数	延べ語数
50 (43%)	3,456
100 (58%)	4,661
400 (83%)	6,671

** 異なり語数

* 単位語生起頻度上位のものから

分布を測定する尺度を求める。第4図は生起分布を示すものであり，この関係を，同一現象を分析した Zipf 法則を利用することで評価を試みた。

2) Zipf 法則の利用

語彙の性格を統計によった手段で導きだした調査としては，Zipf 法則[18]が有名である。これは“$r \cdot f = c$”という関係が多くの言語に見られるとしたものである。すなわち，ある言語表現や，その集団から得られる語彙を調査すると，そこに使用された語の生起頻度（f）とその生起頻度の大きい方から数えた順序（r）との間に“$r \cdot f = c$ (c=const.)”という関係が近似的に成り立つというものである。実例としては英語の文章中の単語の生起での確率は，その順位にほとんど正確に逆比例するものであるとした第5図の関係が有名である。

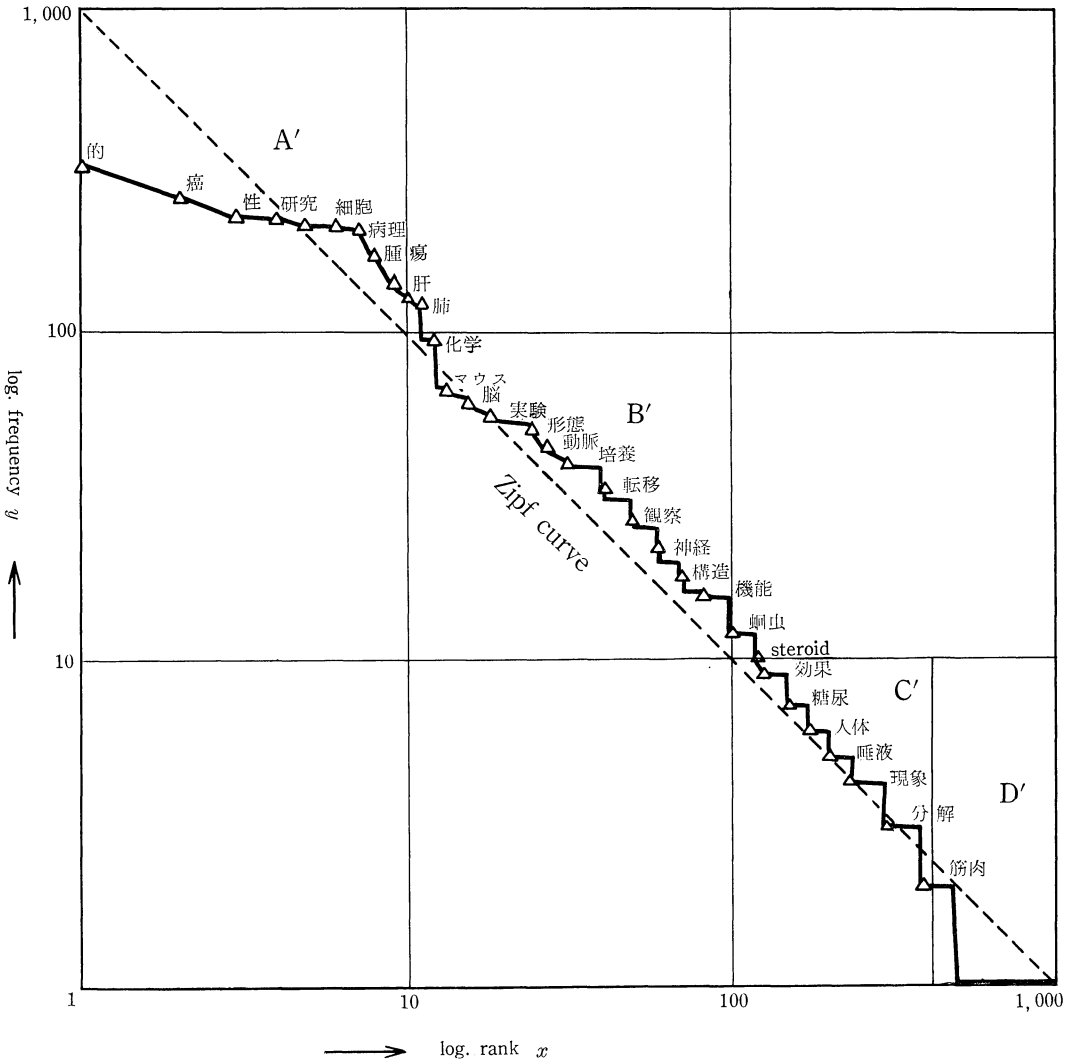
この関係では，例として生起頻度が100番目の単語は，生起頻度が1番目の単語の 1/100 の確率で生起するとしている。Zipf 法則は一種の経験的事実による裏づけによっている。それは生起頻度の大きい順に単語を並べておいて，その n 番目の単語 (w_n) の生起確率 $P_{(n)}$ は次のように表現される。[19]

$$P_{(n)} = \frac{0.1}{n} \qquad \sum_{n=1}^{8727} P_{(n)} > 1$$

この関係がもし厳密に正しいならば，各単位語の生起確率 $P_{(n)}$ と n 番目の単語 (w_n) との関係を描いたグラフは左上から右下への直線になる。しかしここで問題となるのは“$P_{(n)} = \frac{0.1}{n}$”があらゆる単語の状態にあてはまることはないという事である。なぜなら，確率の和は 1 であるから，“$P_1, P_2, \cdots, P_n$”を加算して行くと，P_{8727} まできたとき確率の和が1に達することになり，もし実際にそのとおりであれば，それ以外の単語は生起することはなくなってしまうからである。Zipf 法則はこの点で完全に正しいとは言えない。もしすべての単語の確率が単語の順位に逆比例するならば，すべての単語の確率の和が1よりも大きくなるからである。この点に関して Mandelbrot[20] により理論的に解明されたものがある。しかし，ここで Zipf 法則は次の事実を導いた点に価値がある。

人間の言語行動の1つの特徴を，近似的なものではあるが描写した点である。これについては，Shakespeare と Russele, Eliot の各作品で使用される単語の生起頻度とその順位との関係についての調査でも実証されている。Zipf 法則は，言語はある大きさの基礎語彙を持ち，それらは人間の脳のメカニズムに支配され，コントロールされて用いられるものであることを示す試みと言える。調査では Zipf 法則を一つの公式として測定尺度を求めることにした。標本中での単位語を両対数グラフに

第 4 図　単 位 語 の 生 起 分 布



生起頻度 y とその順位 x とを取り，プロットしたものが第 4 図となる。第 4 図では Zipf の理想直線は，

$\log y = -\log x + \log 10^3$ となる。

第 4 図に示したように生起分布の状態を A, B, C の 3 ブロックに区分してみる。A′ ブロックは順位 1〜10 まで，すなわち生起頻度 368〜100 回まで，B′ ブロックは順位 11〜100 位まで，生起頻度 101〜10 回まで，C′ ブロックでは順位 101〜1,000まで，生起頻度11〜1回までとなる。この各ブロック中で Zipf 直線に最も近似するパターンを示すものは B′ ブロックと C′ の一部分とである。ここで C′ ブロックを近似な部分に限定して残りを D′ ブロックとする。したがって C′ は順位 101〜400 まで，生起頻度 11〜5 回となる。

この B′, C′ の Zipf 直線に近似するパターンを分析すると，先ず Zipf 直線は，

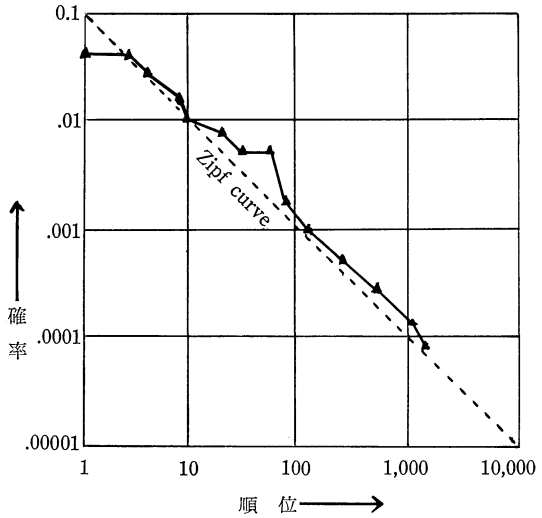
$\log y = -\log x + \log 10^3$ で表わすことができ，

それは　$\log y + \log x = \log 10^3$

すなわち　$\log xy = 3$　であるので

$$y = \frac{10^3}{x} \quad (1 \leq x \leq 2.6)$$ となる。

第 5 図

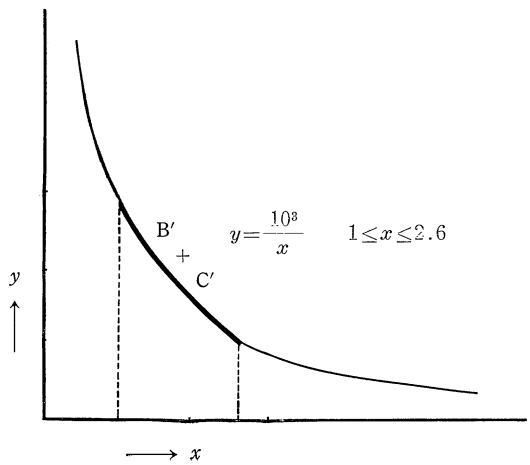


この式は自然目盛では第6図のようにプロットされる。ここでZipf直線に最も近似なパターンを示すB′, C′は順位 10～400, 生起頻度 100～5 回までであることが解る。

そこで生起分布は B′, C′ の部分で最も Zipf 直線に近似な関係を示し, これを自然グラフに変換することで意味ある傾斜を示す。このパターンを語彙量の最も豊かな部分と推定する。

すなわち第6図は生起頻度の極めて高いものは特に注目しなくてもよく, それから生起頻度が低く殆んどフラットになってしまう部分から先も又注目すべきでないことを, したがって中間での B′, C′ が重要であることを意味する。この尺度による測定は後のキーワードの抽出において重要なものとなる。

第 6 図



C) 語彙量の測定に関する考察

①カテゴリー化による尺度は, 一種の意味論的方法を持ち込むもので, 単位語の意味の定義による類の尺度による集合体を造りだそうとしたものである。これにより量による尺度では出来ない測定を試みるもので, 言わゆる語彙のカテゴリーとしての使用量の分析となる。結果はカテゴリー "C>E>A>B>D" のごとく単位語の集合の大小により単位語の類の数と量とを明らかにした。そこで, このカテゴリーを次のような論理集合体で標本の主題概念を表現した。

〝病理学＝{(用途, 性質)×(操作, 処理)×(器官, 器具)×(症状, 現象)×(能因, 対象)}〞

②標本中での単位語の異なり語とその占める割合との相関については函数近似式が成立した。

$$y=19.55\,x^{0.236} \qquad 0<x\leq 1{,}014$$

この式は母集団での単位語の量的尺度となる。第3図での B, C ブロックは全体の 83% を示す。

③Zipf 法則による分析では, 第4図に見られる Zipf の理想直線と比較した場合, 高さの相違は単に標本の大小により決まる問題である。重要なのは曲線の傾斜が標本の大小によらず一定であることである。これは, いかなる言語もある大きさの基礎語彙を持つ事を示すことになる。したがって標本では B′, C′ に近似するパターンがそれに該当する。曲線の右下の部分に見られる階段形は生起頻度の低い単位語が 1～3 回ぐらいで多く生起しているためである。この階段形の部分を Zipf の理想直線のように 45° の傾斜のごとく直線にすると, 一段と傾斜が鋭くなることが解る。すなわち, このことは, 生起頻度の測定では, 1 回の単位語も無視できなく, 重要である事を示す。次に Zipf 法則により推論されることとしては, 次のことがある。コトバはそれを使用する人の頭脳のメカニズム, すなわち能力と言われるものに支配されて特定のパターンを示すものであること, または逆に特定のパターンに自己を適応させることでコトバを選択するものであるということである。このことにより人はだれでも自分の環境の中で, およそ同類の特徴に注目して, 言語活動をしていることによるとも考えられる。この推論を適用すると, 標本については次のことが言える。特に単位語を中心に考えた用語は, 特定主題を記述

する上で一定の数，すなわち語彙量をもつ。この語彙は主題記述上の表現の経済化を目的とするものであるから，日常語に比較して決して豊富なものとはならない。それ故に量よりも質的な範囲が重要なものとなり，そこでつねに一定のパターンに広がる生起分布となる。この現象は標本の大小には関係なく見られるものと推定される。その理由は，用語が日常語のように無意識による統計作用による使用ではなく，特定の主題領域という範囲のコントロール下での意味の働きがあると考えるからである。

④第4図で最も Zipf 直線と近似を示す B′, C′ ブロックの生起分布のパターンは，第6図の“$1 \leq x \leq 2.6$”で成立する“$y=\frac{10^8}{x}$”式で表現される部分となる。

この部分は語彙量に関しては，最も巾が広く豊富であることを示している。この点では第3図における B, C, ブロックすなわち“$10 \leq x \leq 400$ 内で成立する“$y=19.55\,x^{0.236}$”の函数近似式によるパターンでは，単位語は全体の83%を占めている。これらの共通のパターンは，標本の単位語の集合の程度とその状態などの性格を意味する部分となり，いわゆる基本語彙層と考えられる。

D) 確率の利用による尺度

一般に確率は，一つの特定のことがらが出ると思われる程度の表現と言える。ここで，母集団または特定標本中での特定の単位語が生起する事象を w_n とし，その単位語が第 n 回までに n' 回生起した場合 n'/n の n が大きくなっていったときに n'/n がある値にかぎりなく近づくならば，その値を特定の単位語の生起確率とし“$P(w_i)\quad 0 \leq P(w_i) \leq 1$”とする。その時，$P(w_i)$ は，特定の単位語の生起する程度が表現出来る。そこで，種々な単位語が生起する事象をそれぞれ，$w_1, w_2, \cdots w_{n'}$ とするとき，“$P(w_1)+P(w_2)+\cdots P(w_{n'})=1$”が成立すると考えられる。この確率の概念は，母集団での単位語または合成語の生起頻度を示すための尺度として利用される。また，標本中での単位語と単位語との連合，結合，などの組合せによる単位語が同時に出現する場合の尺度として利用される。現存のデータからの確率への変換は，標本中での単位語の使用率，すなわち相対頻度から求められた。(第4表参照)

確率の尺度は，標本中では求めることの出来ない合成語の生起頻度の測定に利用される。合成語の使用率は，標本のように単位語のみの集合では表現することが不可能となる。何故ならば標本は合成語を単位語に再抽出するという認定操作によって求められた単位語の延べ語数の集団にすぎないからである。一方，合成語は2つ以上の単位語の集合によって一つの特定の用語，または句を形成するものであるので，この種の合成語を同一語と認定する基準が明らかでない。このことは分布範囲の設定と認定操作とを不可能とする。そこで合成語の生起頻度は，単位語の生起頻度を確率に変換した $P(c)$ によって，次のように表示する。各単位語の生起確率を $P(w_i)$ とするとき，合成語が生じる確率を $P(c_i)$ とすれば，

$$P(c_i)=P(w_1)\times P(w_2)\times \cdots P(w_n)$$

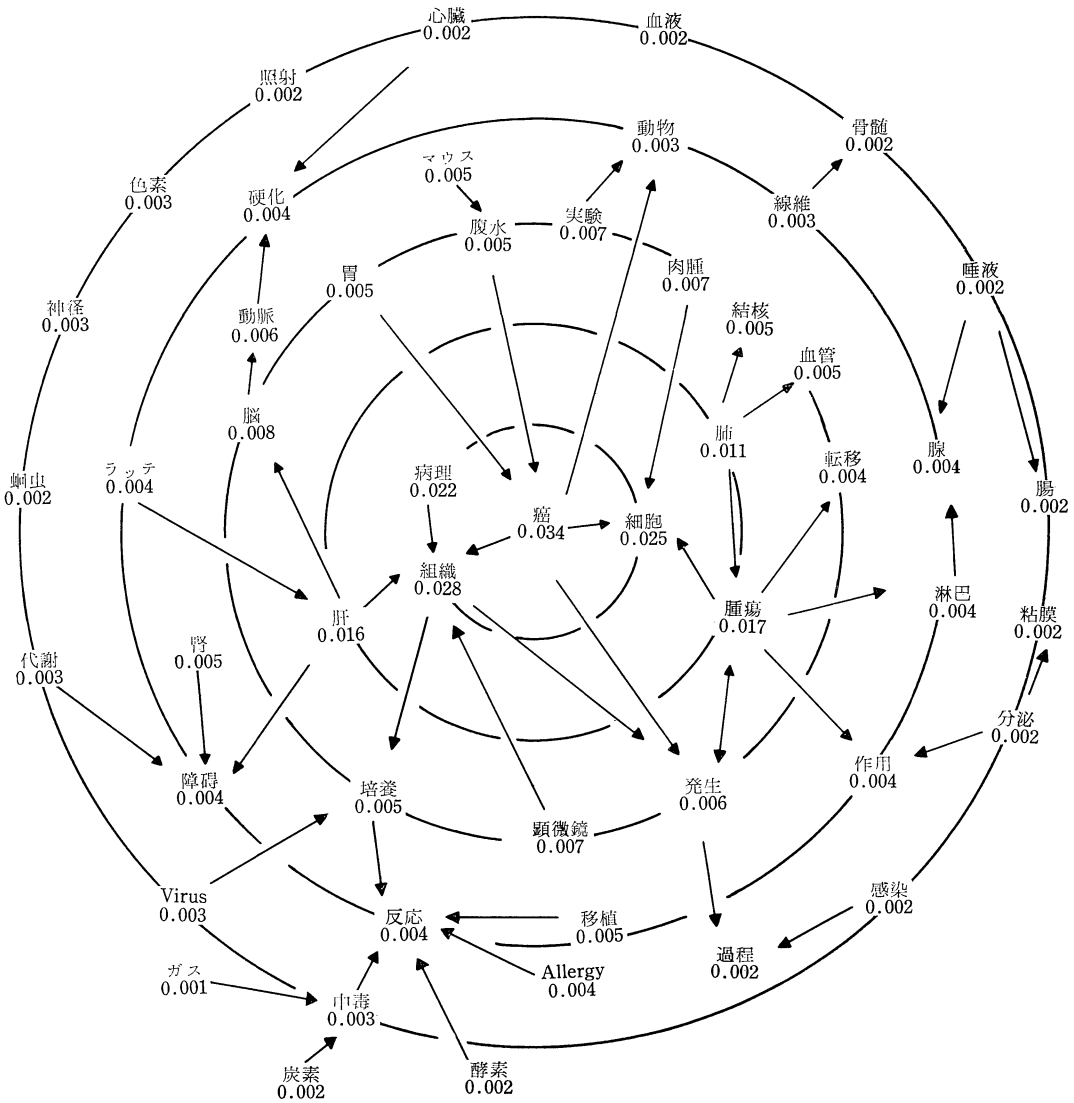
となり，これを共出現率と名付けた。

第 4 表

順位	確率	単位語	順位	確率	単位語
1.	0.029	研究	8.	0.009	化学
2.	0.028	組織	9.	0.008	マウス
3.	0.025	細胞	10.	0.008	顕微鏡
4.	0.022	病理	11.	0.008	脳
5.	0.017	腫瘍	12.	0.007	実験
6.	0.016	肝(臓)	13.	0.007	電子
7.	0.011	肺(臓)	14.	0.007	肉腫

この $P(c_i)$ は，標本中での単位語との集合である合成語の生起する程度を示すものと言える。この種の確率は母集団または標本中で生起する合成語や，その関連語とかが同時に生起する可能性や，その種類を測定するために利用される。それは〝胃〟という単位語があれば〝癌〟がさらに〝粘膜〟が連続して生起するし，また〝細胞〟とくれば〝癌〟や〝培養〟，〝線維〟，〝増殖〟等が連続して生起する程度が高いことを示すことになる。また次のような単位語が，それぞれ確率を持つて生起する関係を示している場合では〝胃 (0.005) → 癌 (0.034) → 細胞 (0.025) → 組織 (0.028) → 培養 (0.005)〟があれば，例として〝癌 (0.034) × 細胞 (0.025) $=8.5\times10^{-5}$〟が示すように，単位語の連続が多い程，生起確率すなわち共出現率が低下することになる。共出現率，すなわち生起確率とその連続する関係を捕える試みとして，第7図に示した単位語の生起確率分布図が作成される。この第7図は生起確率の最も高い単位語〝癌 (0.034)〟を中心に，その高いもの順に周辺に分布する単位語との関係を捕えようとしたものである、第7図では単位語と単位語間の矢印は連続する方向と関係を示すものである。これにより単位語の集合の状態を一目で明らかにしようとしたもので，標本での主題概念の量的構成を把握するための尺度となる。この種の試みはシソーラスの作成への一つのア

第 7 図　単位語の生起確率分布図



プローチにもなると考えられよう。この図での問題点は，単位語の生起確率分布は最も高い単位語〝癌〟を中心にした平面での関係の描写にすぎないことである。

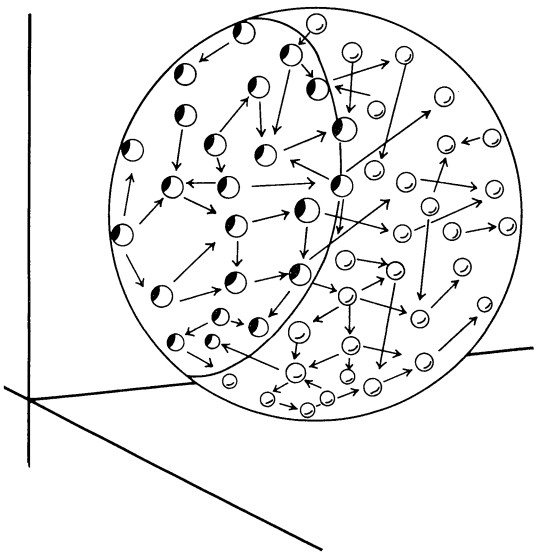
これらの単位語の関係は，厳密には立体面，すなわち球体での関係として捕えられるものと言えよう。第8図は球体での関係を推定しての模型である。単位語を点ととし，その連続する関係を示すものを矢印とすると，それは分子構造式のように可能な限りの方向に可能な手を持つことになる。そこで確率分布の中心は，球体のごとく無数の面にあると言える。したがって一断面が第7図と考えられよう。

V. 索引用語としての単位語の評価

A) 単位語からのキーワードの抽出

キーワードは普通，件名 (subject headings)，記述子 (descriptor) と同意語と考えられる。いずれも，文献の内容を主題分析することで，主題を抽象化したコトバすなわち索引用語と見なされる。ここでは単位語中か

第 8 図　生起確率分布の模型



らキーワードとなるべき単位語群を評価しようとした。そこで標本中の語彙量によって測定される手段で出来るだけ機械的に抽出できる方法を開発してみる。この場合単位語の性格がそうであったように同意同義等の処理，別の表現をすればシソーラス的な処理は，一切行わなれていない点に注意しなければならない。この方法での利点は単位語の数値から判断するということによる，客観的な抽出が出来ることである。この尺度は，次の関係で求めたものとする。

$$k_{(sw)}=\frac{w_{(f)}}{t_{(n)}}$$

$w_{(f)}$：母集団または標本中での w 語の生起頻度

$t_{(n)}$：母集団または標本中での異なり語数

$k_{(sw)}$：母集団または標本中での w 語の使用率

この関係は，たとえば標本中での w 語と x 語との使用率について“$k_{(uw)}>k_{(ux)}$”という関係が成立すれば，その標本中では w 語の方が x 語よりも重要であるという意味を示す。したがって，この方法では生起頻度の高いもの程キーワードとして抽出されることになる。この使用率はリスト 1 で示してある。この傾向を修正するため Edmundson[21] による修正頻度のアイデアがある。それは次の関係でキーワードの抽出尺度を考えたものである。すなわち，

$$S_1=f-r \qquad S_2=\frac{f}{r}$$

$$S_3=\frac{f}{f+r} \qquad S_4=\log\frac{f}{r}$$

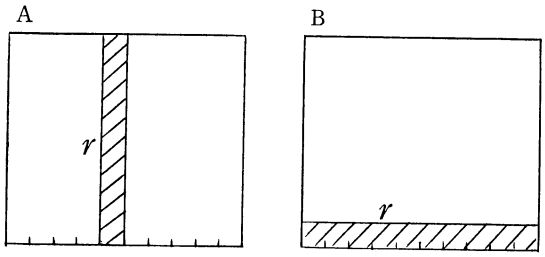
S＝キーワードの尺度

f＝キーワードの対象文献中での生起頻度

r＝キーワードの母集団または標本中での生起頻度

このアイデアは母集団または標本中での生起頻度でキーワードのバラツキを修正するものと言える。しかしこの場合でも，必ずしも充分なものではないとして緒方氏[22]は次の例で考えを述べられている。たとえば r が同じ 100 であっても第 9 図のような極端な場合もあるからである。

第　9　図



この生起分布では S の価には差異は全くない。しかし A と B とでは，B はあらゆる分野に満遍なく生起する常用語となり，A ではそうとは言えない。常用語と言われるものは母集団または標本中で比較的に満遍なく生起するものであるか，または主題内容に関係ないものを言う。それらは生起頻度に関しては，上位にランクされることが当然予想される。例として〝所見，剖検，研究，症例，検討，展望〟等の，それ自体意味的には重要な価値を持たず，論文の形式または慣習的に附加される単位語などや，標本中で生起頻度の上位の語より推定された「癌細胞組織の病理学的研究」の標題が示す主題で当然予想される常用語となる〝癌〟，〝細胞〟，〝病理〟等の単位語である。これらは単位語としては，あまりにも一般的であり，キーワードとはならない。したがってキーワードとしては除外すべきであるが，それによって逆に蓄積，検索上での索引もれ (missing) の原因にもなる。そこで考えられるのは，この種の単位語は合成語としてのキーワードとすべきものとなる。そこから，〝胃癌，腹水癌，肝癌，肺癌，発癌物質，制癌剤，等のレベルでの合成語をキーワードとする考えが導かれる。

調査におけるキーワードの抽出は，先の単位語の語彙量を測定した第 3 図での単位語の異なり語とその占める割合との関係で示された A，B ブロック，さらに C ブ

ロックを加えた単位語400語（83%）を対象として行なった。さらに第4図では，それぞれ A′，B′，C′ ブロックに対応するが，ここで常用語ブロックとして A′ を考えて，これらは除外するかまたは合成語とすることでキーワードとした。

B) キーワードとしての評価

キーワードとしての単位語は言うまでもなく蓄積，検索のためのコトバ，すなわち索引用語となる。これについては，実際の索引用語との対応関係でその価値を評価してみた。現在索引用語集としての医学用語集は日本語のものはないので，英語の MeSH (Medical Subject Headings)[23] との対応を試みた。ここで日本語と英語との対応上での問題点としては，次のことが考慮されなければならない。すなわち，対応させるには二重の障害がある。一つは英語の記述と日本語による記述での表記体の差違と，もう一つは語の示す概念の差違とである。

この点の問題は非常に解決が困難なものとなるが，同時に日本語を英語に翻訳する時に最も重要な問題を提起する。例として，標題での対応関係を次の形で考えてみる。

ⓐ「実験的動脈硬化に関する研究，第一篇，高脂肪食飼育家兎血液脂質及び糖質代謝に及ぼす ATP の影響」では

「Studies on experimental atherosclerosis (1) The effect of ATP (Adenosine Triphosphate) on blood lipid and carbohydrate metabolism in rabbits fed on high fatty diet」

ⓑ「実験的胸膜腫瘍よりの組織培養細胞株樹立とその復元に関する研究」では

「Tissue culture of an experimental pleural tumor and retranslantation of the tissue culture cells」

これらの対応は文脈対文脈となるものであり，単位語のレベルでの対応とは相違するものとなる。本調査では，MeSH 中に単位語に妥当する索引用語がどのくらいあるかで評価の尺度とした。

キーワードとしての単位語は，400語（補助語を除外して382語）に対して完全に適合したものとして193語があった。したがって約51%が，単位語をキーワードとするレベルで MeSH と対応出来ることを示す。第5表は対応例である。なお 193 語の対応での MeSH のカテゴリーでの大小は，Ⓐ—38%，Ⓑ—2%，Ⓒ—31%，Ⓓ—14%，Ⓔ—5%，Ⓕ—1%，Ⓖ—8%，Ⓗ—2%等であり，したがって “A>C>D>G>E>B，H>F” となる。

第 5 表

単位語	索 引 用 語	カテゴリー名
癌(腫瘍)	NEOPLASMS	C
肉 腫	SARCOMA	C
培 養	BREEDING	G
移 植	TRANSPLANTATION	E
色 素	PIGMENT	D
肺	LUNG	A
粘 膜	MUCOUS	A
代 謝	METABOLISMS	G

MeSH のカテゴリーには次のようなものがある。

Ⓐ……Anatomical term（解剖用語）

Ⓑ……Organisms（有機体）

Ⓒ……Diseases（疾患）

Ⓓ……Chemical and Drugs（化学物質及び薬品）

Ⓔ……Technics and Equipment（分析，診断および治療の方法と器具）

Ⓕ……Psychiatry and Psychology（精神医学）

Ⓖ……Biological Sciences（生物科学，基礎医学）

Ⓗ……Physical sciences

以上のカテゴリーにおいて，キーワードとした単位語を評価してみる。評価は，標題用語中でのキーワードの妥当性を分析すること，と同時に，本文の一形式と考える抄録中でのキーワードの状態を分析することで，計量的に求めたものとする。

C) 自然語文中でのキーワード

キーワードとしての単位語が，自然語による文脈でいかなる価値をもつかを評価する。方法としては自然語文脈の形式で記述される論文の抄録を材料とした。この評価の中心はキーワードの性格と，論文抄録と標題との関係すなわち論文標題中のキーワードの価値を求めることにある。この実験での尺度としては検索率という考え方を利用した。検索率（R）とは，特定の文脈での任意のキーワードの生起頻度とその使用率とによって表示されるウェイト（W）と，文脈中でのキーワードの総数（T）との関係

$$R=\frac{W}{T}$$ で表示したものである。

ここで文脈とは次の仮説で捕えたものとする。

すなわち文脈（S）は，キーワード（K）との関係で捕

えると，集合と集合の要素との関係 "$S=(k_1, k_2, k_3, \cdots, k_n)$" となる。ここでキーワード間には，普通の文脈で言う論理的，文法的関係はないものとする。自然語文中では，文脈は次の順序により認定した単位となる。第一に句は単位語（キーワード）の2語以上の集合体で，あるまとまった意味を持つものであること，次に節は2個以上の単位語（キーワード）と句との集合体であること，さらに文脈は2個以上の節の集合体であることになる。そこで暫定的には抄録文中での「～。」で区分される所を文脈の認定単位とする。実験対象としては標本と同じく医学中央雑誌（昭和37年）より任意に20抄録を選定した。次に20抄録から次のデータを引き出す。

ⓐ抄録文の長さ：総語数で示し，漢字，ひらがな，カタカナは各々一字ずつ，外国語と数字は単語または数値を一字とし，また特殊記号，およびそれに準ずるものは除外する。

ⓑ抄録文中での文脈の数を求める。

ⓒ文脈当りの字数を求める。

ⓓ文脈当りの単位語を求める。ただし補助語，外国語は除外する。単位語の認定操作は標題用語から抽出したものに対するのと同じである。

ⓔ文脈当りのキーワード数を求める。

次のような抄録文例では上記のデータは第6表のようになる。

第 6 表

抄　録　例　A	サンプル平均
ⓐ抄録文の長さ……358 字	303 字
ⓑ文脈数（標題）……8 (9)	8.3
ⓒ文脈当りの字数……45 字	38 字
ⓓ文脈当りの単位語……11.5語	14 語
ⓔ文脈当りのキーワード……8.5 語	8 語

＜抄録例 A＞

「吉田肉腫細胞に対するレ線照射の効果の量的分析」吉田肉腫細胞 10^7 個を腹腔内移植されたラッテにレ線 200γ 或は 1000γ を移植72時間後に全身照射した。これら被照射動物の腹水を腫瘍細胞総数と照射に因る細胞学的変化を考慮して検査した。吉田肉腫細胞の増加は 1000γ の照射では48時間以上，200γ の照射では約24時間に亘り阻止され，腫瘍細胞の分裂時間に分裂間期の時間は照射に由り延長した。また腫瘍細胞の異常像は照射後増加したが，これは二つの Peak を示した。異常像の最初の最大値は照射後1時間で分裂細胞に起り，6時間後には分裂間期の細胞に起った。而して第二の最大値は照射24または48時間後に分裂間期の細胞に起った。照射に由る腫瘍細胞の最著明な形態学的変化は分裂細胞での染色体の断裂並に迷錯と，分裂間期の細胞に於ける核破壊とであり，若干の被照射腫瘍細胞は壊死性変化を示した。また分裂後期に於ける腫瘍細胞の照射障碍は分裂前期のそれより強い。

抄録文例 A での認定された延べ単位語数は 104 語で，キーワードは 77 語となった。そこでキーワードを頻度と使用率との積により表示し，ウェイト順にするとリスト3のようになる。

次に抄録文中での文脈（S）の検索（R）率を測定する。

標題（T）では，

T) 吉田肉腫細胞→レ線照射→効果→量的分析

キーワード数……5　　ウェイト……3.75 となり

検索率　$(R)=\dfrac{3.75}{5}=0.75$ となる。

S_1) 吉田肉腫細胞→腹腔内移植→ラッテ→移植時間→全身照射

$R=0.67$

S_2) 被照射動物→腹水→腫瘍細胞→照射→細胞学的変化→考慮→検査

$R=0.97$

S_3) 吉田肉腫細胞→増加→照射→時間→照射→時間→阻止→腫瘍細胞→分裂時間→分裂間期→時間→照射→延長

$R=0.59$

S_4) 腫瘍細胞→異常像→照射後増加→Peak

$R=0.95$

S_5) 異常像→最初→最大値→照射後時間→分裂細胞→時間→分裂間期→細胞

$R=0.62$

S_1) 最大値→照射→時間→分裂→分裂間期→細胞

$R=0.71$

S_7) 照射→腫瘍細胞→最著明→形態学的変化→分裂細胞→染色体→断裂→迷錯→分裂間期→細胞→核破壊→被照射腫瘍細胞→壊死性変化

$R=1.36$

S_8) 分裂後期→腫瘍細胞→照射障碍→分裂前期

$R=0.82$

以上の結果により，R が最も大きいものは S_7 となり，

リ　ス　ト　3

順位	キーワード	ウエイト	頻度	使用率	順位	キーワード	ウエイト	頻度	使用率
1.	細　胞	37.80	15	2.52	13.	染　色	0.37	1	0.37
2.	腫　瘍	10.38	6	1.73	14.	時　間	0.32	8	0.04
3.	照　射	2.47	13	0.19	15.	動　物	0.31	1	0.31
4.	肉　腫	2.10	3	0.70	16.	分　析	0.21	1	0.21
5.	変　化	1.98	3	0.66	17.	検　査	0.12	1	0.12
6.	移　植	0.92	2	0.46	18.	効　果	0.11	1	0.11
7.	形　態	0.60	1	0.60	19.	核	0.09	1	0.09
8.	分　裂	0.50	10	0.05	20.	破　壊	0.07	1	0.07
9.	異　常	0.50	2	0.25	21.	増　加	0.06	2	0.03
10.	腹　水	0.49	1	0.49	22.	阻　止	0.03	1	0.03
11.	ラッテ	0.44	1	0.44	23.	壊　死	0.03	1	0.03
12.	障　碍	0.38	1	0.38					

以下“$S_2>S_4>S_8>T>S_6>S_1>S_5$”となる。$S_7$を自然語文脈に再現すると，

〝照射に由る腫瘍細胞の最著明な形態学的変化は分裂細胞での染色体の断裂並に迷錯と分裂間期の細胞に於ける核破壊とであり，若干の被照射腫瘍細胞は壊死性変化を示した。〞となる。

ここで標題の R は平均 $\overline{R}$ である 0.83 よりも低いことになる。しかし標題が完全な自然語文脈ではない点を考慮すると，この例では決して悪くはないものとなる。実験では抄録中 20 では標題の検索率 R は全体の検索率の平均値 $\overline{R}$ よりは劣っていた。しかしこのことは先にも論じたように，文脈と標題との差違が大きく原因していると考えれば決して低いものとは言えない。そこで以上のことからして論文標題はその内容を比較的に反映していることが推定される。しかしここでの疑問は，先ずサンプルが小さすぎることによる欠点にあると言えよう。

ま　と　め

本稿での調査研究は，現在では，以上で論じた段階までである。今後はこの調査を基礎として新たな方法を開発して行くつもりである。本調査で論じた各点をまとめてみると次のようになる。

標本の採集については，母集団の把握が不明確である。また標本が標題 1,000 であり大きいものではない。したがってそこから抽出した単位語は約 8,000 語であるが，これに基づく結果は部分的考察にすぎない。さらに標本が病理学領域に限定されていること，標本が標題用語に限定されていることにより，用語調査で対象とすべき論文用語やテキスト用語が無視されている。次に，たとえ〝なまの用語〞調査と言っても，対象とした年度，雑誌等により制限が考えられ，それが主題傾向の特色となって現われる。たとえば〝癌〞特にその〝肝癌〞についての特集が多い場合などでは，当然，用語はその関係のものが多くなると言える。

用語の抽出方法では日本語での認定単位が明確化されていないことが挙げられる。そこで極端な場合では，抽出用語の総数における誤差を生じる原因となる。本調査での認定単位も暫定的なものと言える。この問題は，日本語の用語を処理する時つねに直面するものであると考えられよう。用語の分析と評価に関しては，いわゆる計量的処理ができない領域としての〝意味〞がある。この領域の処理には，どうしても semantics, syntax での言語学的方法が必要となる。この研究では，合成語の構造形式を数量化する分析方法についてふれただけであった。

語彙量の測定に関しては，カテゴリー化による尺度により計量面で無視される危険のある単位語の類のバラツキを測定する試みを行なった。しかしこの場合も，シソーラス的操作は一切行なわれてないことにより，単位語の同意，同義，さらに包括，特殊の関係の処理は行なわなかった。統計的アプローチでは，計量言語学での手法から Zipf 法則を尺度として利用した。この経験法則は一つの分析のカギとして利用されたものである。

索引用語（キーワード）としての評価は，単位語から機械的にキーワードを抽出する方法の研究のためであ

る。キーワードは，単位語を全体とするとその部分となるものと考えれば，単に頻度等の数値にのみ依存する抽出尺度には大きな疑問を残すものとなる。やはりキーワードは概念分類法とか件名表に見られるように，一定の制御下にある語でなければならない。そこでキーワードの評価手段としては実際の件名表である MeSH との対応が試みられた。ここでの問題は，日本語と英語との相違の把握である。

自然語文中でのキーワードは，キーワードとして抽出された単位語を標題以下のレベルすなわち抄録文中で捕えようと試みた。抄録文は，標題よりも一層自然語文脈に近い形式で表現されるものである。そこでのキーワードの妥当性と，さらに標題用語の妥当性とを測定する試みを行なった。結果としては標題から判断する医学領域での標題用語は，必ずしも論文内容を忠実に反映していないことが判った。

この研究調査の特色は，計量的または統計的手法に基づいている点にある。この方法では，数値を目安に尺度と測定という関係が重要となる。量的測定が，一般に質的方法に比べてすぐれているという仮説は，量と質の解釈上で客観的側面に限定した点にあると言えよう。質的方法は，量的方法よりも客観的要素の把握の点で明確さを欠いているということになる。それは測定するための特定の尺度を引きだす際の単純化を欠いているということに他ならない。測定とは，多かれ少なかれ，一種の単純化であると解釈される。単純化は，用語の持つ複雑な要素から一つの本質である頻度を抽出する作用と見られる。単純化の過程で重要なことは，〝～とみなす〟ということである。すなわち，一つ一つを詳細にみればすべて異質のものである用語を，集団としての現象として把握してみる場合に，そこに等質のものがあると〝みなす〟という態度で分析することに意味があるのである。しかし，もしもこの種の単純化によって得るものが多ければ問題はないが，一方失われるものが多い場合は重大な危険性があることに注意しなければならない。この点が，統計という単純化操作による利害得失を考慮する上で肝要となる。

（修士課程図書館・情報学専攻在学中）

1) Costello, J. C. "Uniterm indexing principles, problems and solutions," *American documentation*, vol. 12, Jan. 1961, p. 20–6.
2) Bernier, C. L. "Indexing process evaluation," *American documentation*, vol. 16, Oct. 1965, p. 323–8.
3) Bourne, C. P. *Methods of information handling*. New York, Wiley, 1963. 241 p.
4) Gillum, T. L. "Compiling a technical thesaurus," *Journal of chemical documentation*, vol. 4, no. 1, 1964. p. 29–32.
5) N. L. M. *MEDLARS index manual*. 1966.
6) 津田良成．〝MEDLARS の日本におよぼす影響，〟**情報管理**, vol. 8, no. 2, 1965. 2, p. 17–20.
7) 斉藤　孝．〝医学用語の統計的調査（1）病理学領域における論文標題用語を例として，〟**きたさと**, vol. 5, no. 2, 1966. 9, p. 35–87.
8) 森　茂樹．病理学総論．金原出版，1953．639 p.
9) 医学中央雑誌，1962，vol. 172–182.
10) Luhn, H. P. "Keyword-in-context index for technical literature," *American documentation*, vol. 11, Oct. 1960, p. 288–95.
11) Kraft. D. H. "A comparison of keyword-in-context indexing of titles with a subject heading classification," *American documentation*. vol. 15, Jan. 1964, p. 48–52.
12) Montgomery, C. "Machine-like indexing by people," *American documentation*, vol. 13, Oct. 1962, p. 359–66.
13) 緒方良彦．〝統計的操作による自動抄録法，〟**情報管理**, vol. 8, no. 9, 1965. 9, p. 3–12.
14) 奥津敬一郎，成田正雄，〝日本文法への数学の応用，〟**数理科学**, vol. 2, no. 2, 1964. 2, p. 36–9.
15) Vickery, B. C. *Classification and indexing in science*, 2 d ed. Academic Press, 1959. p. 19–48.
16) Walton, T. S. A formal indexing language for automated document retrieval systems. <A.D.I. *Automation and scientific communication*. Washington, D. C., 1963. Pt. 1> p. 21–2.
17) 増山元三．実験公式の求め方．竹内書店，1962．（現代応用数学双書）
18) 渡辺　修．〝言語学における情報理論，〟**数理科学**, vol. 3, no. 5, 1965. 5, p. 28–32.
19) 細井　勉．〝暗号と字引――直線法則とジップの法則――，〟**数理科学**, vol. 4, no. 8, 1966. 8, p. 41–6.
20) Pierce, J. R. 鎮目恭夫訳．サイバネティックスへの認識；情報理論とその展望．〔*Symbols, signals* and *noise*〕白揚社，1963. 366 p.
21) Edmundson, H. P. and Wyllys, R. E. "Automatic abstracting and indexing—Survey and recommendations," *Communications of the Association for Computing Machinery*, vol. 4, no. 5, 1961, p. 226–34.
22) 緒方良彦．〝自動抄録法の問題点，〟*Library sci-*

ence, no. 4, 1966, p. 17–27.
23) N. L. M. *Medical subject headings*. 1967.

参 考 文 献

藤川正信．語の連合と概念構成——Coordinate indexing 方式の再検討——＜第3回ドキュメンテーション研究集会発表論文集．1966＞p. 99–104.

水谷静夫．〝統計的自動抄録法の問題点，〟**計量国語学**, no. 27, 1963, p. 1–13.

山川邦雄．〝文献分類の数量化の考察，〟**数理科学**, vol. 3, no. 5, 1965. 5, p.58–62.

国立国語研究所．〝現代語の語彙調査——総合雑誌の用語——前，後編，〟**国立国語研究所報告**, 12–13, 1957.