Library Science No. 4 1966

自動抄録法の問題点

Problems Related to Automatic Abstracting Methods

緒方良彦
Yoshihiko Ogata

Résumé

Through the observation of the results of two experiments on preparation of automatic abstracts between July 1964 and February 1966, the author tried to consider problems of techniques employed to automatic preparation of abstracts of Japanese texts.

As for the selection of key words, besides counting raw frequency of words occurrence, we tried to test the method proposed by Edmundson-Wyllys, namely "relative frequency" approach, when applied to Japanese texts. The "relative frequency" is used to adjust occurrence of words in a target document by conditions of occurrence of words in document population which includes the target document. This method is meaningful to the selection of key words for automatic indexing but can't be applied to the selection of key words for automatic abstracting. Based upon this interpretation, we tried to select key words employing a method to adjust frequency of words occurrence in individual document by conditions of occurrence of words in the document itself.

In the process of extracting a sentence used for constructing element of abstract, there are two methods, namely to consider each document as a unit and to consider each chapter and paragraph as a unit. The author discussed merit and demerit of those two methods.

In the automatic processing of Japanese texts, the automatic recognition of processing unit is always the center of debate. In this article, the author introduced a method to recognize automatically processing unit required for the selection of key words in automatic abstracting and indexing only, and gave evaluation of the results.

(Institute of Behavioral Sciences. Division of Automatic Processing of Language and Document)

自動抄録法の問題点

I. はしがき

「抄録」とは，いうまでもなく，文献内容の要点を簡潔に表現する目的で，原文献を量的に縮少して記述しなおす操作をいうので，「自動」抄録とは，そのような操作の主体を電子計算機が演じることを意味している。つまり，電子計算機を用いて，文献の抄録をつくらせるために必要な技術が自動抄録法である。

ところで，現在行なわれている自動抄録法には，大別して2つの流儀が存在する。第1は，情報の表現に用いられている言語の意味とか構造に着目して，その面から抄録をつくっていこうとする方法で，第2は，言語の内部に立入らないで，外面的な現象自体を手がかりとして抄録をつくらせる方法である。ここでいう自動抄録法は，これらの方法のうちの後者に限ったものである。従って，文献内での「ことば」の使われ方だけに注目し，その構文的ないし意味的側面には全く触れないで抄録文を作成しようとする態度を一貫してとっている。このような方法を，ここでは一応「統計的自動抄録法」と名づけておこう。

さて，これまでに，実験もふくめて行なわれてきた統計的自動抄録法は，周知のごとく，原文献中から何らかの基準にもとづいて所定数の文を抽出してきて，それらを文献にあらわれている順序に書き並べて抄録文としている。通常の抄録では，抄録者が自身の「ことば」で文献の要点を書き綴るわけであるが，統計的自動抄録法では，原文献の中にあらわれる生の文を，文 (sentence) 単位で抽出する，といった手続きを前提としている。従って，このような意味での自動抄録法においては，原文献の中からどのような文を抽出すれば，原文献の要点が最もよく投影できるか，が問題となる。ここでは，そのような観点から，具体的な実験から得られた問題点を詳述しようとするものである。

この報文中で各種の実験データを使用しているが，これは，1964年7月から1966年2月にかけて，2回にわたって行なった実験から得られたものである。これらの実験は，外務省電子計算機室よりの委託にもとづいて，財団法人計量計画研究所が行なったもので，その概要は次のとおりである。[1, 2, 3]

第1回実験

対象文献

社会科学文献（文献A）

1964年7月に行なわれたイタリアのモロ首相の，同国議会での施政方針演説の要旨で，在イタリア大使館より外務省あてに送られてきた公式の報告文。

科学技術文献（文献B）

「東芝レビュー」に掲載された「大容量タービン発電機の新技術」と題する論文で，主として大容量化に伴う冷却技術の最近の進歩および東芝の新技術を解説したもの。

文学作品（文献C）

Washington Irving の作品「Sketch Book」中の1篇「John Bull」。イギリス人の気性の一端をジョン・ブルという人物に託してまとめたエッセイで，その日本語訳を使用。

抄録技術

19通りの技術を用いて自動抄録文を作成した。

評価法

人間による文抽出抄録（文献A，Cは10人，文献Bは8人）を作成し，人間による抽出パターンと19通りの自動抽出パターンとを比較し，人間抄録からの距離の大小によって優劣をきめる方法をとった。

第2回実験

対象文献

社会科学文献（文献D）

「世界週報」所載の「米中激突の方向へ――ベトナムをめぐる国際関係」と題する論文で，ベトナム紛争の政治的解決の前途が暗いものであることを解説したもの。

抄録技術

28通りの技術を用いて自動抄録文を作成した。

評価法

本実験参加者が会議によって，それぞれ8文(535字)および17文(1,086字)からなる2種類の文抽出による抄録文を作成し，28通りの自動抽出の各文をこれらと比較した。

第1表は，実験に使用した4文献の定量的性格をまとめたものである。

なお，これらの研究・実験に直接参加したのは以下のとおりである。

古郡廷治（計量計画研究所）
大村睦子（計量計画研究所）
緒方良彦（計量計画研究所）

第 1 表

文　献	A	B	C	D*
章区分	21	8	なし	9
文数	174	240	190	124
延べ語数(a)	6,715	7,326	6,221	8,198
異り語数(b)	1,425	1,216	1,642	
1 語の平均使用度数 $\frac{a}{b}$	4.71	6.02	3.79	
1 文の長さ(平均延べ語数)	38.5	30.2	32.7	66.1

* 文献Dにおいては語数勘定が不可能であるので，語数はすべて字数単位になっている。

高橋達郎　(日本科学技術情報センター)
藤川正信　(慶応義塾大学)
笹森勝之助　(日本科学技術情報センター)
また，その際，統計数理上の専門的な助言を，次の2氏より受けたことを付記しておく。
水野欽司　(計量計画研究所)
山川邦雄　(株式会社日本ビジネス・コンサルタント)

II. 人間抄録文の特性

人間による原文献からの文抽出による抄録文を作成し，その抄録文に何らかの統計的な特性が発見できれば，延いてはそれが統計的自動抄録法の技術につながることとなるわけである。そこで，2 回の実験において，各文献に対して，第2表のような人間抄録文（文抽出によって）を作成した。

第 2 表

	抄録者数	1 人当り目標抽出文数
文献 A	10 人	20 文
文献 B	8 人	24 文
文献 C	10 人	20 文
文献 D	8 文 (535 字) よりなる抄録文 17 文 (1,086字) よりなる抄録文	

これら延べ30種の人間による文抽出抄録は，結局は，自動抄録の各手法によってつくられた抄録文の評価に当って，その基準となるものである。

ところで，これら30種の抄録文の外面的な特徴を探ってみよう。ここでは，抄録文の論旨の把握の仕方や，そのたどり方には触れないことにする。その点を加味した問題については，文抽出による抄録文の有効性という観点から，後述する予定である。

抄録文として抽出された文が，抽出されなかった文との間にみせる主要な特徴の第1は，1 文当りの平均延べ語(もしくは字)数である。例えば，文献Aについてみると，全体として(174 文)の 1 文当り平均延べ語数は 38.5 であるが，6 人以上が一致して抽出した10文の平均延べ語数は 62.6 語と極めて大きくなっていて，この傾向は他の文献についてもみられる。このことから，抽出文数に指定があるとき，抄録者はやはり長い文を抽出しがちになることが明らかである。このことは，抄録者の個人別にみても以下のとおり，全文平均よりかなり大きい値を示していた。

第 3 表　文献Aに関する個人別平均延べ語数

抄録者	1 文当り平均延べ語数
A	54.9 語
B	55.3
C	38.1
D	47.3
E	51.5
F	47.1
G	63.5
H	33.0
I	48.7
J	51.3
全文 (174 文)	38.5 語

□ は平均以下。

次に，抽出された文が，各パラグラフのどこに位置しているかを調べてみると次のようになる。（第 1 回実験分のみ）

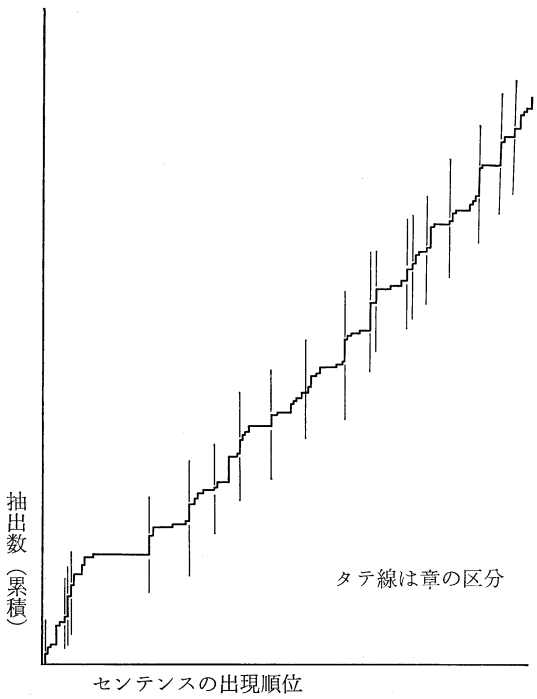
第 4 表

文　献	A	B	C
全パラグラフ数 (a)	54	117	26
抽出された第 1 文数 (b)	42	60	22
b/a	0.74	0.52	0.85

これで明らかなように，抽出文は各パラグラフの第 1 文である割合が極めて高い。また，これとは逆に，各パラ

グラフの末尾文を抽出した例は極めて稀であった。殊に文献Dに対する抄録（合議によって作成した）では，文献の最後尾のパラグラフと2文以下で構成されているパラグラフを除き，他のパラグラフの末尾文が抽出されている例は皆無であった。これらのことから，パラグラフの前半の部分には，文献の要点に重要な関連をもった特殊な文が含まれている可能性が極めて強い，ということが明確にいえそうである。

第1図は，文献Aについての全抄録者が抽出した文を，原文献での文の出現順に累積していったものである。こ



第 1 図　文献Aに関する累積抽出文数

れによると，各章（図ではタテ線で示されている）の初めで累積数が急激に上昇し，あといくつかの追加がでるが，章の後半は概して平板になっている。この傾向から，次のようなことが明らかとなる。第1には，文献自体に章節の区分が設けられている場合，抄録者は各章（または節）から満遍なく文を抽出する傾向があって，例えば，章節の長さなどの相違（文献Aではかなりの差異がみられる）によって，章別にある重みを与えるようなことは殆んどなく，むしろ，各章の内容を平均的に満遍なく見わたせるような抄録文を作成するようにみられること。第2には，文献自体には，やはり文献の主題・内容に直接に関連をもった部分と，そうでない部分（重要でない，という意味ではない）とがあり，つまり，各部分が均質でなく，抄録者はそれらの関係の深い部分（一連の文の系列からなる）に先ず注目し，そのような部分から抽出すべき文を選んでいるらしいことである。そして，更にいえることは，そのような関連の深い部分の認識には，抄録者間に余り個人差が生じないが，そのような部分の中で，どの文を抽出するかでは，個人差が現われる，ということである。

以上のような顕著な特性があるにもかかわらず，原文献の10分の1ないし20分の1という範囲の文数を特定できるような統計的特性を，人間抽出文から導き出すことはついに不可能であった。ということは，結局は統計的自動抄録法に対する悲観的な見通しを与えるものであることは否定できない。しかし，このことが直ちに統計的自動抄録法の実用上の価値まで，全く否定してしまうことにならないのも，また当然である。いま仮りに，上述のような悲観的見通しを打破するために，文法的ないし意味論的な自動技術を導き入れたとして，現段階では，この面からのみのアプローチもまた，統計的手法と同じような迷路に陥ることは必定で，結局は，これら各種の方式のそれぞれの特質を生かした併用的な方向こそが，最も実際的な方法だと考えられるのである。従って，統計的自動抄録法に対しても，一層の努力を傾注する必要が，依然として存在することになる。

III. キーワードの選定

統計的操作のみで抄録を作成する場合，次のような仮説に立脚していることは周知のとおりである。

（1）　ある文献の内容もしくは主題に深い関係をもった「ことば」は，その文献の中でしばしば繰返して用いられている。

（2）　そのような，しばしば繰返して用いられている「ことば」を数多く含んでいる文 (sentence) は，その文献の内容もしくは主題を，最もよく表わしている。

この仮説は，かの Luhn の着想である[4] が，かれは，このような仮説にたって，抄録文作成の手順を次のようにすすめた。

第1段階　　当該文献に使用されていることばのすべてについて，いわゆる語彙調査を行ない，各語の出現度数をしらべて，度数表をつくる。

第2段階　　この度数表から，いわゆる機能語(func-

tion word) と常用語 (common word), および予め定められた度数以下の語を取り去り，残ったものをキーワードする。

第3段階　各文ごとに，その中に含まれているキーワードの延べ出現度数を調べ，あらかじめ定められている方法で，各文に評点を与える。

第4段階　評点のたかい文から順に，予め定められた数に達するまで順次文を抜き出し，それらを文献中にあらわれている順序に書きならべる。

これらの手順は，基本的にはLuhn以外の他の統計的手法の殆んども踏襲しているものであるが，このような操作手順を前提とする限り，統計的手法の問題は，結局は

（1） キーワードの選定法

（2） 文に対する評点法

の2点に尽きる。そこで，以下，それらの点に絞って考察をすすめよう。

キーワードを選定する際の基準に関しては，いくつかの観点が考えられる。

まず，選定の際の対象の範囲を，どのひろがりを単位として選定をすすめるかということから，

a 1　文献の全体をひとつのまとまりとして，一括対象とする。

a 2　章節(もしくは，それに相当する)区分が予め設けてあれば，その区分毎に，それぞれをひとつの独立した対象とする。

ことが考えられる。

(a1)は，対象となる1個の文献について，語彙調査を行ない，(a2)は各章毎に別々に語彙調査を行なう，という操作上の差異が出てくる。ひとつの文献がいくつかの主題をふくんでいるとき，その中のどれか特定の主題のみに直接関連の深い語だけが，かたよってキーワードになる，という危険は充分予想されるところで，そのための配慮として，(a2)の方法がとられることになる。文献内で自動的に認識できる適当な区分が，予め著者等によって設けられていれば，それらの各区分を独立した対象とみて，それから均等（もしくは，区分の長さその他を加味した割合で）にキーワードを選定していけば，文献全体に満遍なく関連をもったキーワードのセットが出来あがる，というわけである。

例えば，文献Aは，21の章がそれぞれ独立して経済，政治，社会，労働，外交などの分野を専門的に取り扱っていて，他章の内容とは一応関連なしに記述されている。このような文献の性格をも反映して，文献Aに関する度数表の上位にある語についてその散らばり方をみると，176 語中の 70% に近い 118 語が，明らかに特定の章で顕著に出現度数が高くなっていた。したがって，文献全体から一括して選定されたキーワードのセットが，ある特定分野に密接に関連した語だけから成り立つ，という可能性は充分考えられるわけで，仮りに，このような特定分野に偏向したキーワードのセットができた場合に予想されることは，これらを手がかりとして抽出されてきた文は，やはり，その分野に関連のある内容を含んでいる可能性が強い，ということである。つまり，一定数の文を抽出する場合，偏向をもったキーワードのセットを手がかりとするか，普遍的なキーワードのセットを手がかりとするかによって，当然抽出されてくる1組の文の内容は全体として変化することになる。さらに，このことを一歩すすめると，指示的抄録と報知的抄録とのいずれかを，自在につくれる可能性につながる，と考えられるが，いまはその点には触れないでおこう。

キーワード選定の第2の観点は，語彙調査の結果判明した全異り語のなかから，どのような条件をもった語をキーワードとするか，逆にいうと，どのような語を篩い落とすか，という問題である。

Luhn は英文に関して，例えば冠詞，代名詞，関係代名詞，接続詞などを含めた常用語 (common word) のすべてと，所定の出現度数に満たない語を取り除くことにしている。なお，Luhn の場合，common word として予めリストを作成しておく方法をとり，いわゆる文法的な機能語 (function word) と，ある部門で普遍的に使用される常用語 (common word) とを特に区別しておらず，その両者を common word list 中で扱っている。

さて，一言にしていえば，キーワードは文献の主題・内容を特徴づけるようなものでなければならない。そこで，Luhn は，語彙調査の結果，出現度数順にならべられたリストから，主題・内容そのものに関係のない常用語(かれの場合は機能語を含めている)と，度数の少ない語を取りのぞいた。ただ，ここで問題となるのは，常用語の範囲をどのように定めるか，という問題がある。

常用語というのは，対象となっている文献が属している母集団内において，比較的に満遍なくどの文献にもあらわれるような語である。そのような語は，たとえある文献で出現度数が上位を占めているからといって，必ずしもその文献の内容を有効に示さないことがある。従って，そのような常用語はキーワードの選定からはずそう，

というのである。つまり，文献から直接に得られた度数のみによって，単純にキーワードを選定していくのでなく，母集団における使用状態をも考慮して，有効なキーワードを選び出そうとするわけである。

ところで，この常用語の具体的な扱い方については，Luhn のように，予め不要語のリストに登録しておいて，それらリストの中に含められている語は，どの文献についても篩い落としてしまう，というやり方が第1に考えられる。言うまでもなく，常用語の内容は，主題分野によって実際には変化すべきものであるから，主題分野毎に別々の不要語リストを準備しておけばよい。

また，Edmundson ら[5]は，キーワードの条件として，対象文献中での出現度数は高いが，その属する文献母集団内での出現度数の低い語が，文献の内容を最もよく特徴づけるものであるとして，次のような4種の測度を提案した。

$$S_1=f-r$$

$$S_2=\frac{f}{r}$$

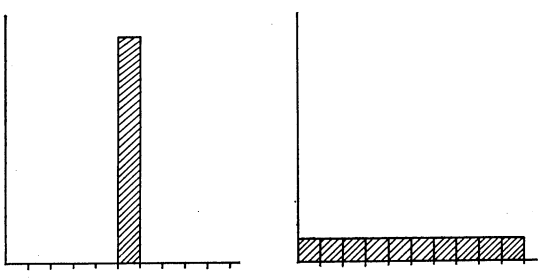
$$S_3=\frac{f}{f+r}$$

$$S_4=\log\frac{f}{r}$$

S は語の重要度
f はある語の対象文献での出現度数
r はある語の文献母集団内での出現度数

いずれも，f を r によって修正しようとするもので，Edmundson らは“relative frequency”と称している。

この“relative frequency”は，母集団における総体的な出現度数のみをもって修正しようとするもので，必ずしも充分な配慮であるとはいい難い。例えば，r が同じ100であっても，次のような極端な場合がおこり得る。第2図における，A，Bはともに $r=100$ であって，この限りでは S の価に差異は生じない。しかし，この場合，AとBとでは内容を特徴づける能力には大きな差があると見るべきであろう。つまり，Bはあらゆる分野に満遍なく使用されている，ということで，まさに常用語であるが，Aはその名に値いしない，極めて偏った用い方がされている。



第 2 図

あることばが，特定の主題・内容を指示するのに，どれくらい有効かということは，その文献母集団内でどれくらい頻繁に使用されているかということだけでは充分ではない。同時に，母集団内に内在している主題別の層の間に，どのような散らばり方で使用されているかをも，考慮に入れなければならないはずである。

そこで，このような点を考慮に入れた新しい測度を第1回実験では採用してみた。その測度は，国立国語研究所が行なった現代雑誌の用語用字調査において用いられた「語の基本度」という考えを借用したものである。[6] この基本度というのは，特殊分野にだけよく使われるのでなく，概して広く満遍なく使われるような語を「基本語彙」と呼ぶとすると，ある語を「基本語彙に属せしめるか否かを決するてがかりとなり，かつ，一次元の尺度で表わせる」ような量をさしている。基本度 g は，次式によりあらわされる。

$$g=\phi(p,\ Sc)$$

ここで，p はある語の使用率
Sc はある語の散らばり度
ϕ は基本度函数

なお，同研究所では，実験式として次式を算出している。

$$Z=-0.6356+1.5825x-0.4181y$$

ここで，
$$x=\log_{10}p+5$$
$$y=\log_{10}Sc+3$$
$$Z=0.01g$$

ところで，このような文献の母集団内における語の現われ方によって度数を修正していく方法には，基本的な疑問がある。自動索引法[7]においては，文献の母集団内において，個々の情報を内容的に識別する必要がある。つまり，個々の概念間の関係を予め階層的に位置づけておいて，これらの概念を指示する語もしくは記号を当該情報に附与していくことになる。従って，そこでは，母集団内における語（もしくは記号）の相対的な関係がつねに問題となる。自動索引におけるキーワードは，個々の情報の主題・内容を特定する能力が必要であると同時に，他のキーワードとの相対的関係をも特定できるもの

でなければならない。このような観点にたつとき，自動索引におけるキーワード選定に，対象文献外の，母集団におけるその語の使われ方が重要なカギになることは疑問の余地がない。

だがしかし，自動抄録法においては，いささか問題の性質が違ってくる。ここでは，ある文献の主題内容が把握でき，それを最もよく反映できる1組の文を特定できればよいのである。それらの文が，他の文献とどのような相互関係をもつかは，直接的には何んら考慮される必要がない。自動抄録（単に抄録といってもよい）では，あくまでも，ひとつの文献だけが問題である。そこで，自動抄録におけるキーワードの選定にあたっては，母集団における語の使われ方を考慮するのでなく，むしろ，対象文献中での語の出現状態のみを忠実に反映できるような基準を設けることの方が重要だ，ということになる。

このような観点にたって，第2回目の実験に当っては，新しい測度を採用してみた。それは次のような根拠に立った測度である。

文献を構成している文の系列にランダムにあるキーワードを所定の数だけ配当する試行を行ない，そのキーワードが配当された文（1であらわす）と配当されなかった文（0であらわす）との文系中でのあらわれ方を調べてみると，1の系列と0の系列とに2分できる。

そして，ある特定の文に着目したとき，1があらわれるか0があらわれるかの確率は，

$$\frac{m}{m+n}$$ （1があらわれる確率）……(1)

または，

$$\frac{n}{m+n}$$ （0があらわれる確率）……(2)

によってあらわすことができる。ここで，mは例えば1の文献全体での出現度数，nは0の出現度数をあらわす。このような(1)もしくは(2)式に与えられるような確率に従って，1もしくは0の系列が出現すると考えてよい。しかも，その出現は正規分布に従うとみなすことができよう。

そこで，文献中における出現度数の大きい方から順次，語の個々につき，次式(3)～(5)によりεの値を算出して，有意水準5%(もしくは10%)で片側検定を行う。この際，棄却されるということは，(1)または(2)式の確率からみて，特異な系列をあらわした語であった，ということを意味しているわけである。従って，そのような特異な出現を示した語をキーワードとすることにして，棄却された語（つまり，ここではキーワードと認められた語）が，予め定められている数に達するまで，順次検定をすすめていく。

$$E_{(u)}=1+\frac{2mn}{m+n} \quad \cdots\cdots(3)$$

$$V_{(u)}=\frac{2mn(2mn-m-n)}{(m+n)^2(m+n-1)} \quad \cdots\cdots(4)$$

(1)と(2)より

$$\varepsilon=\frac{U-E_{(u)}}{\sqrt{V_{(u)}}} \quad \cdots\cdots(5)$$

ここで，m, nは，それぞれ1もしくは0の出現度数，Uは0と1の各系列の数の和

以下，文献Dに対する3種類のキーワードをリストにしておく。

〔A〕

ベトナム
アメリカ
主　義
ソ　連
中　共
共　産
ベトコン
解　決
政　権
会　議
戦　争
政　府
政　治
中　立
政治交渉
国　際
平　和
三　月
前提条件
軍　事

〔B〕

ソ　連 (4)
政　権 (9)
会　議 (10)
中　立 (14)
モスクワ (23)
首　相 (26)
革　命 (28)
人　民 (29)
デ　モ (31)
軍　隊 (35)
闘　争 (39)
ウイルソン (43)
抗　議 (44)
関　係 (45)
援　助 (46)
ホー・チミン (50)
カンボジア (52)
シアヌーク (55)
大　使 (57)
訪　問 (58)

〔C〕

1. ベトナム　アメリカ　戦争　軍事
2. ベトナム　主義　共産　ベトコン
3. ベトナム　アメリカ　主義　政権　ホー・チミン(50)
4. ベトナム　アメリカ　主義　共産　戦争　平和
5. ベトナム　解決　会議　中立
6. ベトナム　アメリカ　ベトコン　前提条件

7.	ベトナム	ソ連	中共	会議
	政治交渉	三月	外相(59)	首相(26)
	訪問(58)	モスクワ(23)		
8.	ソ連	政府	デモ(31)	抗議(44)
9.	ベトナム	ソ連	中共	政権　政治

〔A〕は，文献D内における各語の出現度数の大きい方から20位までを，度数順にならべたもの。

〔B〕は，出現度数の大きい方から，上記5%検定を行って棄却された語。( )内の数字は，出現度数の順位。

〔C〕章毎に(9章からなる)，出現度数を調べ，大きいものから全体として異り語が20語になるようにとったもの。ただし，他章との重複，同度数語などの関係で結局，異り語としては26語になっている。( )内の数字は，〔B〕と同様，文献内での出現度数の順位。

IV. 抽出文の選定

ある基準にもとづいて選定された1組のキーワードを手がかりとして，抄録文を構成すべき文の選定に入るわけであるが，その際の選定基準をどのように設定すべきかは，先にも述べたとおり，キーワード選定の問題とともに，統計的自動抄録法の2つの柱となる。そこで，この問題を若干考察しておく必要がある。

抽出文の選定を行なう際，予め選定の対象から除外しておいても差支えないような文があれば，後の処理がそれだけ簡略化されることになる。そのような立場から，人間が原文献から抽出してつくった抄録の特徴が役立つ。先に見たとおり，人間の抽出傾向のうち，きわだった特徴として，各パラグラフの前半から抽出してくること，ならびに，各パラグラフの末尾文を抽出することが殆んどないこと，などがあった。そこで，第2回実験においては，

(1) 1文当りの延べ字数が，全文の平均延べ字数の2分の1に満たない文

(2) 3文以上から構成されているパラグラフの末尾文

を抽出対象から予め除外することとした。ただし，(1)，(2)とも，最終パラグラフには適用することを避けた。文献D（全文数124文）に関してみると，これらの適用によって除外された文が25文あったが，これら25文中には，人間が抄録文として抽出した文は皆無であった。

さて，かくして残った文に対して，ある基準をもととした評価を行なうことになるが，評価に当っては，2つの観点からの問題がおこる。第1は，評価の対象範囲の問題で，第2は，各文に対する評価の測度の問題である。

対象の範囲に関しては，普通，

(b1) 全文をひとつの対象とみる。

(b2) 章節（もしくはそれに相当する区分）のそれぞれを独立の1単位とみなして，それぞれの区分を独立した対象とする。

のいずれかが，考えられるところである。両者の優劣は一概には論じがたく，前にキーワード選定の際の対象範囲に関して述べたような，種々の考慮を必要とする。そして，結局は，この問題は，キーワード選定の対象範囲との関連において論じなければ，無意味である。

Luhnは，キーワードが潜在的にもつ抽出文判定能力(resolving power)は，判定すべき対象の語数の増加につれて，逆に減少していくことを指摘し，文章を細分していって，その細分化された各部において高い評点を得た文を抽出することが有効である，旨をのべている。[4] また，水谷静夫氏も，日本文についての机上実験で，章節ごとに選定したキーワードのセットを用いて，文献全体から抽出すべき文を選ぶ方法〔(a2)と(b1)の組合せ〕が好成績であった，と報告している。[8] われわれの2回の実験結果からは，それらのことを裏付ける明確なデーターは得られなかった。しかし，一般的に言って，次のことが言えそうである。

(a1—b1) 報知的性格が強い抄録

(a1—b2) 方法的に余り意味がない

(a2—b1) 最も偏向性の少ない無難な抄録

(a2—b2) 指示的性格が強い抄録

次に，文評定の測度に関しては，すでにいくつもの提案がなされている。最も単純なものとしては，2回の実験で用いた，

(c1) 1文中に現われるキーワードの延べ出現度数

(c2) 1文中でのキーワード密度

がある。(c2)に関しては，1文中に現われるキーワードの延べ出現度数 (K) を，その文の延べ語数もしくは延べ字数 (N) で割った値 (K/N) をとるのが普通のようであるが，キーワードの集中している文を強調するため，(K^2/N) の値をも使ってみた。

この他に，有名なLuhnの“Cluster”というのがある。各文について，両端をキーワードにはさまれ，そ

の両キーワード間に非キーワードが3語以内介在しているような語系列を，Cluster と呼び，各 Cluster の重要度 (s) を，

$$s=\frac{p^2}{q}$$

ここで，p は Cluster 内のキーワード数
q は Cluster 内の非キーワード数

の式により算出して，これをもって文の評点とした。1文中に2個以上の Cluster があるときは，s の大なる方の値をもって，その文の評点とする。

このような Luhn の態度は，文中に出現するキーワードの出現度数のみでなく，それらの文中での相対的な位置をも重要度に加味しようとしたもので，そのような態度は，Edmundson らによっても，引継がれている。[5]

日本文について，この Luhn の“Cluster”方式が通用できるかどうかは，興味ある問題で，第2回実験の際，人間による抽出文と，原文から任意に抽出した（文系列の等間隔抽出）同数の文とについて，キーワードの文中における位置関係を調査してみたが，両者に顕著な差異が見当らず，現在のところ，“Cluster”方式の適用には，やはり問題があるようである。その理由のひとつとして，日本文における主語と述語との分離があげられそうであるが，確実なところは，現段階ではまだ明言できない。

測度の問題に関連して，キーワードに対するウエイトづけの問題がある。通常は，各キーワードに対してはウエイトづけを行なわず，同一ウエイトとして処理しているが，これには矢張り疑問の余地が残る。Edmundson らも，その必要性をしきりに説いており，例えば，タイトル，パラグラフの第1文，サマリーの部分にある語に対しては，ウエイトづけを行なって，出現度数を修正すべきことを述べている。

文献Dの書名，副書名，章見出しに現われる語について検討したところでは，それらの中にあらわれる語の，本文中における出現度数が極めて少ない，ということが明らかとなった。しかし，さらに人間による抽出文との関連でみると，両者が指示している概念としては極めて近接していて，シソーラスその他による語の統制を実行する必要が認められた。

V. 日本文の操作単位の認識

自動抄録の処理を実行するとき，操作単位をどのように機械に認識せしめるかは，処理の前提をなす問題である。いわゆる分かち書きをしない日本文を対象としたとき，処理の手がかりとなる操作単位が，自動的に認識できないとなると，それ以後の操作がいかに効果的にはたらいても，その実用的意味は，大いに割引く必要がある。そこで，ここでは，自動抄録法そのものではないが，操作単位の自動認識の手法について簡単にふれておく。

いわゆる言語情報の自動処理と呼ばれる範囲に入る主なものとして，

（1） 自動抄録法
（2） 自動索引法
（3） 自動分類法
（4） 機械翻訳
（5） 自動内容分析

などがあげられる。このうち，(4) と (5) とは，その処理からして，文法的にも正確な分かち書きをしなければならない。他方，(1)，(2)，(3) などは，文法的なアプローチをとれば別であるが，統計的なアプローチをとる限り，キーワードの選定さえ出来れば，必ずしもその他の部分を精密に分かち書きする必要はない。そこで，ここでは，自動索引や自動抄録の処理上必要なキーワードの選定を可能とするような，その限りでの操作単位の自動的認識の手法を紹介しておこう。従って，文法的な正確さを要求されるような処理目的には，このままでは応用できないものであることを，予め付言しておきたい。

さて，ここで述べる手法には，一つの前提条件がある。それは，原文献が漢字まじりの普通の日本文であることを必要とする。つまり，全文カナ文字書きやローマ字書きのものには，このままでは通用しない。それらはそれで，また処理可能な手法があるが，ここでは，漢字まじりの日本文である場合についてだけ述べる。

以下は，操作手順である。

〔第1手順〕

句読点，引用符等の切れ目として自動的に識別できる部分で切る。

〔第2手順〕

次いで，各系列について，ヒラガナの系列とヒラガナ以外の文字の系列に2分し，ヒラガナの系列と，ヒラガナ以外の文字の系列のうち1字で構成されている系列を捨てる。ここで残った系列を暫定単位と称する。

(例) アンダーラインは暫定単位

第2次モロ内閣／は／経済危機克服対策／を／優先的／に／進／め／，／かつ／長期的／には／経済社会組織面／の／構造的諸改革／を／実行／するための／中道左派政治／を／行／う／。

(解説)

ここで，1字のものを捨てたのは，第1回実験の結果からも明らかとなったが，1字構造のキーワードが極めてまれにしか現われず，それを無視しても最終結果には殆んど影響がない，と考えたからである。従って，この手法をとる限り，1字構成のキーワードは皆無となる。

〔第3手順〕

各暫定単位について，カタカナ，数字，ローマ字のいずれか1種類の文字のみの2字以上の連続が含まれている場合には，その部分のみは最終単位とし，それによって区切られた前後の各系列は，それぞれを1つの暫定単位とする。その際，1字のみからなる暫定単位は捨てる。

(例)

第2次モロ内閣

第2次 → 暫定単位
モ　ロ → 最終単位
内　閣 → 暫定単位

純イデオロギー問題

純　　→ 捨てる
イデオロギー → 最終単位
問　題 → 暫定単位

〔第4手順〕

各暫定単位について，各文字を頭字とするあらゆる文字の組合せをつくる。但し逆順，飛越し，1文字は除く。ここで出来た組合せを「組合せ単位」と称する。

(例)「構造的諸改革」の組合せ単位

構造　構造的　構造的諸　構造的諸改
構造的諸改革　造的　造的諸
造的諸改　造的諸改革　的諸　的諸改
的諸改革　諸改　諸改革　改革

〔第5手順〕

全組合せ単位について出現度数調査を行ない，各組合せ単位毎に，出現度数（f）と構成文字数（l）の積（lf）を算出する。この（lf）の値を修正度数と称する。

〔第6手順〕

全組合せ単位を，それぞれ同一派生源の暫定単位毎にグループ化する。

〔第7手順〕

同一派生源に属する全組合せ単位について，修正度数を比較し，その最大のものをもつ組合せ単位を最終単位とする。もし，最大の修正度数をもつものが2箇以上あるときは，構成文字数（l）の大きいものを，（l）も等しい場合は，暫定単位の文字の配列順序の後方の文字を頭字にもつものを最終単位とする。ここで出来た最終単位により分断された，暫定単位の残りの部分は，それぞれを新たな暫定単位とする。その際，1字構成のものは捨てる。

(例)「中央準備委員会」

組合せ単位	l	f	lf
中央	2	5	10
中央準	3	2	6
中央準備	4	2	8
中央準備委員会	7	2	14
央準	2	2	4
○ 準備	2	8	16
準備委員会	5	2	10
委員	2	4	8
委員会	3	3	9
員会	2	3	6

(解説)

最終単位をなるべく短かい小分割の方向へもっていく方が，最終的キーワードとして望ましければ，修正度数を用いないで，単純な度数（f）の値で比較すればよい。こうすれば，文字数の小さい組合せ単位が有利になる傾向がでる。しかし，余り小分割を強めると，例えば

イタリア／共産／党

のように，「共産党」が2分されてしまうような可能性が大きくなる。そこで，長い文字系列のものを有利にするように（lf）の値で比較する方法をとってみたわけである。

〔第8手順〕

各暫定単位について，最終単位のいずれかと完全に一致するものがあれば，それを最終単位とする。暫定単位の1部に最終単位を含んでいる場合には，その部分は最終単位とし，残りの部分は暫定単位となる。最終単位と一致しないし，その一部にも含んでいない暫定単位は，それを最終単位とする。

〔第9手順〕

第8手順によって新たに出来た暫定単位につき，第8手順を繰返し，新たな暫定単位が出来なくなったとき，操作を終了する。

この操作手順を用いて第2回目の実験では，文献AとDとについてのキーワードを自動的につくりあげた。いま，文献Aに関するキーワードの1部を紹介しよう。

ちなみに，第1回目の実験では，文献Aに対しては，人間が次の簡単なルールを用いて，分かち書きを行ない，機械で語彙調査を行なったのち，キーワードを選定した。

〔分かち書きルール〕

（1） 文を単語で切る。（文節から付属語を切り離したものを単語とした）

（2） 接頭語，接尾語を切り離す。

（3） 数をあらわすものは，ひとつの単位とみなす。

（4） 3字以上の漢字は，慣用的にみて，各々がまとまった概念を表わしていれば，2つ以上に分ける。

（5） サ変動詞は，語幹と語尾を切り離す。ただし，「関する」と「対する」等語幹が漢字1字の場合は切り離さない。

さて，文献Aを使って上記操作手順を適用して得た結果を，度数順に大きいものから列挙すると，以下のとおりとなった。(20位まで)

1	経　済	4	生　産
2	計　画	5	政　府
3	増　加	6	政　策
7	所　得	14	物　価
8	1963	15	目　的
9	内　閣	16	必　要
10	投　資	17	手　段
11	労　働	18	安　定
12	措　置	19	予　算
13	問　題	20	通　貨

この結果を，人間の分かち書きにもとづいて作った第1回実験の際の度数表と比較してみると，順位に若干の上下があることと，度数に若干の差があることを除いて，本質的には殆んど同じ結果であった。従って，このことから，これらの語を手がかりとして自動索引や自動抄録のためのキーワードを選定していくことに，殆んど実用上の問題はない，と考えてよいだろう。

（計量計画研究所）

1) 外務省電子計算機室．統計的自動抄録法．1965. 3. 2 vols.

2) 緒方良彦．“統計的操作による自動抄録法，”**月刊 JICST 情報管理**，vol. 8, no. 9, 1965. 9, p. 3-12.

3) 緒方良彦，古郡廷治，木村睦子，高橋達郎．統計的自動抄録法．＜日本科学技術情報センター．第2回ドキュメンテーション研究集会発表論文集．東京，1965＞ p. 219-24.

4) Luhn, H. P. “Automatic creation of literature abstracts,” *IBM journal of research and development*, vol. 2, no. 2, Apr. 1958, p. 159-65.

5) Edmundson, H. P., and Wyllys, R. E. “Automatic abstracting and indexing; survey and recommendations,” *Communications of American computing machinery*, vol. 4, no. 5, May 1961, p. 226-34.

6) 国立国語研究所．現代雑誌九十種の用語用字　第3分冊：分析．東京，1964. p. 7-22.

7) ここでいう自動索引 (Automatic indexing) とは，情報を検索する際の手がかり (key) となるTag（個々の情報の内容を識別するための目じるし）を自動的に決定する操作を意味している．

8) 水谷静夫．“統計的自動抄録法，”**計量国語学**，27号，1963.12, p. 1-13.