Library and Information Science No. 15 1977

自動索引の動向と逆説的アプローチ

Trends in Automatic Indexing and Its Paradoxical Approach

加藤徳義
Noriyoshi Kato

Résumé

Beginning with Luhn's pioneering works, many investigators have been involved in the development of theories and techniques of automatic indexing. Practicality of basic researches, however, is discussed rather pessimistically, because the present trend shows a direction toward computer-aided indexing in rather restricted environment.

Recognizing the fact that the reason for the difficulty lies basically in the inherent complexity of natural languages, this paper suggests to take a paradoxical approach to the problem of automatic indexing; practical systems can be established in special environments where automatic indexing systems may not be hampered by such language complexity.

Typical pathological reporting systems, now operational in the U.S., in which findings are given orally on the spot and processed automatically are discussed as a good example of feasible quasi-automatic indexing systems from the following points: communication language consists of highly specialized and well-defined non-phrases; the volume of data to be processed is sizable; and the format is rather simple.

加藤徳義：日本アイ・ビー・エム株式会社
Noriyoshi Kato, IBM Japan, Ltd.

自動索引の動向と逆説的アプローチ

I. はじめに

機械翻訳の研究が，これまでの莫大な研究投資にもかかわらず，その対象とする自然語の神秘の前に挫折を余儀なくされていると同様に，自動索引の研究も決定的なゆきづまりをみせていることは，実に残念である。

一方，情報量の増大はいまや，人力による索引能力の限界をおびやかすに十分すぎるほどであり，情報処理の現場での自動化への要望は高まる一方である。こうした緊急性は，自動索引に対して，より現実的方向を求めている。

本論文では，現状までの文献批評をつうじて，今後の自動索引のとるべき，具体的方策の考察を試みた。ここで提起されるのは，仮に，自動索引研究の進展をはばむような困難性を，その対象とする自然語がもたないとすれば，実用的な自動索引システムは存在しうるのではないかという逆説的発想である。処理対象を，図書館資料からはなれて，病院内部資料のひとつである病理学報告書にうつし，この逆説的仮説を展開してみた。病理学報告書の入力処理は，基本的に索引作業であるにもかかわらず，これまでの自動索引研究者たちから話題とされたこともないが，上に述べた仮説の検定の場としては恰好の自然語環境を備えている。

II. 自動索引法

A. 自動索引法の概観

情報量の増大に対応する，情報の効果的検索方法や，圧縮技術の開拓が必要とされているのは，いうまでもないことである。情報の圧縮については，物理的・形態的圧縮と主題的·内容的圧縮の両面を考えることができる。例えば，ハードコピーで存在する技術報告書をマイクロ・フィッシェに撮って保存するという行為は，情報の物理的・形態的圧縮とみることができる。また，ある雑誌記事を抄録するという作業は，その原論文を主題的・内容的に圧縮するものである。ただし，圧縮変換後も，原論文の情報の総量を全く減ずることなく，抄録という短い表現に変換することはできない。さらに，件名標目を考えるとき，原情報を内容的につきつめて，ひとつの標目を与えるという意味から，同様に内容的圧縮と認められる。この場合も，件名標目は自然語に比べて，極めて単純な構文体系をもち，冗長度も極めて低いので，圧縮と同時に失われる情報も多い。

種々の内容的圧縮には，共通して，検索という目的がある。この意味で，情報の内容的圧縮は，その圧縮度の相違を問わず，主題索引作業（以下，本論文では，とくに限定しないかぎり，索引は主題索引を意味する）ととらえるのが妥当である。

索引作業は，Meadow が指摘するように，〝語彙と構文とからなる索引言語を用いて，ファイル中に蓄積されたデータ内容を代表すること〟[1] である。この観点にたてば，抄録，索引，分類などの行為を，索引作業として，一本化してとらえることができる。

Baxendale によれば，一般に索引には次の 3 つの機能がある。

1. コレクション中に含まれる情報に対して濃縮されたキーを与えること。
2. 文献の著者とそれを探索する者との意味的差異に橋渡しすること。
3. コレクション中の文献を区別するための道具となること。[2]

抄録は，その主たる目的が上記の 1 の機能を果たすことである。また，分類の場合は，上記 3 の機能を主目的とする。このように，各索引言語は，それぞれ上記機能のひとつ以上に力点がおかれている。索引言語のそれぞれによって，索引語の呼称もまちまちであるが，基本的に同じ文脈でとらえられるので，ここでは一般的に「索引語」を用い，個々のシステム例ではそれぞれの用語（例えば，ディスクリプタなど）を尊重して使用する。

さて，情報検索システムの経済コストの大部分が，入力処理段階に費やされることはよく知られているが，この点で索引を含めた入力処理面でのコンピュータ導入による経済的貢献が期待されている。

人間による索引作成には，ふたつの明白な問題点が指摘される。ひとつは，複数の索引作成者間で生じる索引

結果のばらつきであり,他のひとつは,ひとりの索引作成者が,精神的状況の相違,経験による索引作業の質的成長などによって,異なる索引結果をもたらすことである。

これに対して,コンピュータ・プログラムとしての自動索引システムは,ある設定されたアルゴリズムのもとでは,そのアルゴリズムが変更されないかぎり,時間の経過に関わりなく,同一文献には常に同一索引語を付する。自動索引はこのような均質性の点で,人間による索引に比べて高い信頼性を保障する。

それでは,索引の自動化とは一体具体的にどんなことなのであろうか。仮に,完全な自動索引システムが存在するとすれば,そのシステムはべすての手作業による索引作業過程を自動的に行なう。つまり,そのような理想的システムは,自動的に原文献を読みこみ,内容を自動的に分析し,文献のファイルにおける自然語の自動的分析により生成された索引言語を用いて,自動的に索引語を与え,さらに,自動的に質問を分析して,適合文献を検索する。

ここで,基本的に2つの制約がある。第1に,自然言語はもともと動的性質をもち,ある新しい文献がコレクションに入れられる場合,そこには全く予期できない語彙や意味が含まれている場合がある。したがって,文献に書かれた自然語を完全に分析し,定義できるとは保証できない。第2に,現在の技術では,人間へのインターフェイスとしての印刷された,または手書きの文献を機械がそのまま読解する段階にはほど遠い。入力準備作業としての機械可読形態への変換の必要性は,自動索引の実用化にとって障害となることは否定できない。

したがって,自動索引とは,機械可読文献を自動的に分析し,自動的に索引語を割当てることと定義するのが実際的である。

索引語を選定する際に,文献にあらわれた語をそのまま用いる自動索引システムは,自動抽出索引とよばれる。KWIC (Keyword-In-Context) 索引はその典型的な例である。これに対して,自動割当索引は,索引語として統制語彙を用いる。語彙統制機能は通常,辞書またはシソーラスによって果たされる。索引言語の特性は,システム全体のパフォーマンスに強く影響を及ぼすので,こうした語彙統制装置の作成方法は,自動・非自動の別を問わず,自動索引の諸技術と深く関係する。

抄録や分類が基本的に索引と同じ文脈でとらえられるという立場から,自動索引は自動抄録や自動分類の研究から学ぶものが多い。自動抄録は,実際のところ,文中に内容を表現する語を探し,何らかの得点計算基準にしたがって,自動的に文を選択・抽出するものである。この場合,語の重要度を反映すべく,語の重みづけ (term weighting) が使われることが多い。また,自動分類との関連では,カテゴリー名をある文献に自動的に割当てるという作業は,その文献をカテゴリーに分類するのと同義である。

自動索引のパフォーマンス評価という面では,自己の自動索引システムを MEDLARS (Medical Literature Analysis and Retrieval System) の人間による索引結果と比較して述べた,Salton の主張に注目したい。

> 全自動の文章分析・探索システムは,伝統的なマニュアルの文献索引システムで得た結果に比べて劣るような検索パフォーマンスを形成しているようには見受けられない。マニュアルの索引および探索式設定は,その索引作成者や探索者が蓄積コレクションと利用者の要求を完全に理解している場合には,例外的に素晴らしい結果を導き得る一方,このような条件に合致しない時には,はなはだ好ましからざる探索結果も出し得る。反対に自動的システムは,その網羅的な入力データと複雑な分析方法とにより,極端に悪い結果を示すことはめったになく,また,しばしば全く満足のいく検索行動を作り出す。[3]

Salton の主張に代表されるように,自動索引のパフォーマンスは,通常,マニュアルのものとの比較で評価されることが多い。しかし,Salton の比較研究を含めて,一般に比較のための資料が十分な量でなく,個々の比較研究が各々別個のコレクションを対象としていることがほとんどであるため,一様に論ずるには無理があるといわざるを得ない。

この点では,1961年の Borko の研究[4] のように,同一のコレクションに対する,別々の自動索引システムの比較検討が必要であろう。

B. 自動索引の試み

1960年代初頭,にわかに華やかになった自動索引の試みのひきがねになったのは,1950年代末,IBM の Luhn が発表した一連の研究[5] であった。Luhn の文章分析法は,基本的に,文献にあらわれた語の統計的特質,つまり,語の出現頻度に依っている。さらに彼は,基本技法をすすめて,語のクラスター化,分野による語彙の相違,相対頻度,共出現頻度などを考慮した方法について示唆を与えた。

Luhn の自動索引への貢献はもともと，重要出現頻度語による文章選定手続きからひきだされたものである。Luhn はもっとも基本的な測度として，文献 i におけるターム k の出現頻度 $f_i{}^k$ とターム k のコレクション中の総出現頻度 F^k を考え，次のように定義した。

$$F^k=\sum_{i=1}^{n} f_i{}^k$$

ただし，n はコレクション中の文献数

この式は，のちのあらゆる自動索引システムの基礎となった。

入力上の困難をさけるため，Luhn は文献のフルテキストのかわりに標題のみを使用し，KWIC 索引を作りだした。Luhn の索引技法は，索引語の人為的な統制なしに，著者の用語に基づいて，文献内容を機械的に操作できるという可能性を示したものとして評価される。したがって，この技法は，前記の自動抽出索引と認められる。

Luhn の自動抽出索引に対して，1958年，Swanson が示したのは自動割当索引の原型ともいうべきものであった。[6] Swanson の実験では，自動化は単に，予め分析された手がかり語 (clue words) に原文を機械で照合するという方法がとられた。ここでは，ある件名標目にいくつかの手がかり語を用意し，文献中にそのどれかがみつかれば，単純に対応する件名標目を割当てた。したがって，この方式はユニターム索引に代表される伝統的なマニュアル索引方式の単なる機械化であったと考えられる。

Maron は，こうした伝統的索引手法をとらず，確率論的手法を導入した。[7] Maron の立場では，コレクション全体を知ることは不可能であり，こうした不確定な領域では，あるディスクリプタが発見されたという事実は，限定された確率においてのみ，対応する主題カテゴリーに割当てられる可能性があるとされた。

これに対して，Borko は因子分析法を応用し，手がかり語と文献の相関関係を出発点とした方法を試みた。[8),9)] 方法論的には Borko の実験はさほど評価されないにせよ，前述のように他の自動割当索引法（具体的にはMaronのもの）と同一サンプルで比較した点で注目すべきである。さらに，この比較の結果，一般には因子分析法より確率論的手法が自動索引に有効と考えられるようになった。

この他，Borko の方法に類似した，Stiles[10] の研究や，Baker[11] の潜在構造分析，Williams[12] の差別係数による索引語選定のための閾値設定方法の提案などが示されたが，1960年代はじめのこれらの研究は，いずれも後の研究に強い影響を与えなかった。

1965年，Damerou[13] は，ある文献に含まれる総語数と同じ数のランダムな語の集合を考えたとき，ある特定の語が当該文献に実際にあらわれているより，ランダムな標本中に多くあらわれない確率を語の重要度の基準としようとした。この確率分布は幾何級数的分布として正確に計算できるとされている。いま標本が十分大きく，また，全体の中のそれぞれの語が十分少ない割合で分布しているならば，二項分布あるいはポアソン分布でほぼ近似できる。ある文献をとりあげたとき，語の重要度ランクをつけるために，このポアソン標準偏差を用いようという提案であった。このような観点は，近年，Bookstein と Swanson の語の出現分布の確率論的モデル[14] に受け継がれた。彼らの仮説によれば，同一文献中に集中しがちな語は索引語として有効である。あるコレクションで索引語となりうるような語と，そうでない，または内容を表現しないような語との区別は，無作為性からの逸脱度を示す，なんらかの統計的測度によって行なわれる。仮に，非重要語が多数の文献の中に無作為に分布しているとして，また，その分布型を数学的に表現できるならば，重要語は逆に，非重要語の分布を表わす数学的な分布型が，ある語の実際の分布を説明するのにどれだけ不適当かを測定することによって選定できる。

Bookstein と Swanson は，重要語とは，その語の示す内容がコレクション中で扱われる程度からみて，文献のクラスを区別するような語を考え，すべての重要語について2つの文献のクラスを定義するのに，ちょうど2つの扱い方があるとして，これをモデル化するために，2ポアソン・モデルを適用した。つまり，語 w が1文献中に k 回出現する確率 $P(k)$ は，

$$P(k)=\pi\frac{e^{-\lambda_1}\cdot\lambda_1{}^k}{k!}+(1-\pi)\frac{e^{-\lambda_2}\cdot\lambda_2{}^k}{k!}$$

（ただし，λ_1, λ_2 はそれぞれのクラスにおける語の平均出現数，π は一方のクラスにその文献が属する確率）であらわされる。

Harter[15),16)] はこのモデルを基礎として，文献 d における，語 w の索引性 (indexability) の測度，または，相対重要度 β を導いた。

$$\beta=P(d\cdot I/k)+Z$$

ここで，Zはそれぞれのクラスの重複度に基づく効果の測度で，次のように定義される。

$$Z=\frac{\lambda_1-\lambda_2}{\lambda_1+\lambda_2}$$

この β 測度は，人間による索引で割当てられた索引語の単純な文献内出現頻度より常に優れた結果を示したとしている。β 測度は，特定コレクション用辞書作成のための，十分な索引性を備えた語彙の選定に有効な道具と考えてよいであろう。

語の出現事実を計量的に測ることによってキーワードを選定しようとする統計的分析手法では，特定の文脈における，語の品詞種類を識別したり，特定の意味を決定したりすることはむずかしい。そこで，言語研究の分野で各様に研究されている構文分析や意味分析の諸理論を自動索引に応用しようとする動きがみられるのである。

Artandi による，プロジェクト MEDICO[17),18)] は英文薬学情報の自動索引の実験で，とくに，リンクの自動生成に特徴をもつ。1つの文章中に共出現する，2つ以上の語は，その文献中の同一文脈に同時に属する，との仮説によって，それらの語の特定の連結関係を自動的に示し，あいまい性を減少しようとするものである。この生成実験の結果，リンクの適正率は約70%であったが，さらに興味深い点として，リンクで示された語と語の距離について，適正と判断された組合わせの平均が 3.71 語であったのに対し，不適正と判断されたものは 7.08 語という結果が明らかにされた。[19)] また，本実験では，入力データとして，フル・テキストの場合と抄録の場合を比較して，文体が長くなるにつれて，リンク生成の適合率が低下すると指摘された。[20)] 一般に，抄録文の各文はフル・テキストより長く，このため，語間距離の長い，つまり，不適正な確率が高い組合わせを多く生成すると考えられよう。

Lockheed Palo Alto 研究所では，9年間にわたる自動言語解析の研究が行なわれ，自動索引への構文分析法の応用実験も試みられた。[21),22)] 同研究では主として語法の分析により，品詞の決定や意味用法を識別するなどの可能性が試みられたが，なかでも，'word government' とよばれる，各語の文法的，意味論的統治法の概念を採用したことが注目される。この統治法の採用により，特定の語について，品詞の型，用法の型，意味，他の語との関係が整理された，ある種の表を用いて，次のような構文的または意味論的分析がかなり可能であるとしている。[23)]

構文分析

1．名詞句の領域設定
2．前置詞句修飾の決定
3．不定詞の定義
4．名詞・動詞の区別のあいまい性解消
5．分詞用法のあいまい性解消

意味分析

1．統治語[24)] の意味上のあいまい性解消
2．前置詞の意味上のあいまい性解消
3．主語・目的語，述語動詞修飾語の役割のあいまい性解消

この他，構文構造的関係を索引法にとりいれたものとして，Hillman[25)] の研究や，Syntol プロジェクト[26)] があげられる。

これまで概観したような，統計的，構文構造的，意味論的諸手法の集約として位置づけられるのが SMART (Salton's Magical Automatic Retriever of Text) システム[27),28)] である。この自動文献処理システムでは，主として蓄積過程で構文分析的手法が，また検索過程で統計的手法が用意され，各々の諸手法は処理オプションとして，比較検討が可能になっている。文献中の語は，語幹シソーラスにより，語幹と語尾に分離され，各語幹は概念コードと構文コードのペアに置きかえられる。語尾辞書との対照によって，語尾にも構文コードが付される。フレーズ辞書および句構造辞書は，それぞれ，名詞句・前置詞句等の識別，および文の木構造分解による句概念の識別が行なわれる。こうした詳細な文章解析をとり入れた索引システム例は，現在のところ他にみられないが，パフォーマンスの点では，他のシステムとの比較検討が十分ではなく，安易な結論は避けなければならない。

III. 自動索引法の困難性と方向転換

A. コンピュータ補助索引

概観したように，これまでかなりの数の実験が行なわれてはいるものの，1960年代初期の試験的な諸研究と比べて，最近10年間の研究に基礎的な進歩を認め難いのは事実である。このため，完全な自動索引システムは実現不可能なのではないか，といった悲観的結論を急ぐむきも多い。

人間の知的作業をコンピュータ・プログラムによって置きかえようとする研究は，一般に，そのあまりの複雑

さ，難解さのために，マン・マシン・システムを指向する傾向があるが，自動索引も決して例外ではなく，人間の知力を借りた，あるいは，単にコンピュータが人間の作業に部分的に介入する，といった方向の研究が盛んになりつつある。

マン・マシン・システムとしての半自動，またはコンピュータ補助索引は，索引作業そのものを人間の知的活動にあると認め，そのうえで，コンピュータをどのように利用していくかを研究するものと受けとれよう。

Bernier によれば，索引におけるはじめての進歩のひとつは，

……コンピュータが，タームの表示，または標準索引語への翻訳をするなどして，主として辞書編さん者の役割を果たすことにより，索引者を助ける，コンピュータ補助索引であろう。いわゆる自然語索引作業は，辞書編さんのプロセスと索引のそれとを，たいへんうまく分離し，コンピュータに著者と索引作成者の使ったタームを探させ，さらに標準索引語に翻訳させる。コンピュータに登録された語彙にみつからない新しいタームは，索引作成者にではなく，それらを同義語であるとか，むしろ既設の上位語の下位概念であると判定してシステムに記入できる人間の辞書編さん者の手に委ねる。この種のコンピュータ補助索引は魅力的である。[29]

さらに，Fangmeyer は，次のようにコンピュータ補助索引の長所を列挙した。

1．抄録や文献中にあらわれた専門的概念の相対的重要度を区別できる。
2．すべての文献にアクセスできる。
3．その文献だけでなく，参考図書や専門家または他の適当と思われる情報源にまで，適切な索引をするための援助を求められる。
4．文献に明示はされていないが含蓄された概念を，定式化し索引するための，帰納的推理を適用することができる。（割当索引）
5．探索要求の分析，探索ストラテジーの定式化，さらに探索スクリーニングに参加することにより，システムの利用者の要求に慣れ親しむことができる。[30]

コンピュータ補助索引の典型は，DDC (Defence Documentation Center) の MAI (Machine-Aided Indexing) システム[31]にみられる。このシステムは1967年以来開発を続けられ，年間約1千万語にのぼる文献の索引に使用されている。

MAI では，英単語ひとつずつをエントリーとする“Recognition Dictionary”と，システムが許容できる構文型式の辞書である，“Format Dictionary”が用いられる。前者は，不用語リストを兼ね，不用語とならない語は，この辞書によってその品詞種類を認識される。この認識のレベルは，形容詞，独立して索引語となる名詞，他の非不用語と結びついて索引語となりうる名詞等で，登録された語はただひとつの対応するコードを割りふられる。品詞種類を示したこのようなコードの列，例えば，形容詞＋弱性名詞の“interactive retrieval”は，型式辞書に照会されて，その構文型式が正当か否かを検定される。こうして出力されるのが，候補索引語であり，索引者はこの候補索引語のリストから妥当な索引語を決定する。

このように，MAI システムは部分的構文分析を採用し，2語以上の候補索引語を提示でき，人間の索引作業の補助を行なっていることが評価される。ただし，単語各々にただひとつの品詞種類を与えるという方法は柔軟性に欠け，候補索引語の抽出もれを残す。MAI システムに似た例は，Ohio 州の Dayton University における材料科学分野資料の索引システム[32]にもみられる。

コンピュータ補助索引がいつから生まれてきたかの文献的判断は困難であるが，ここでは，Doyle の意味地図[33]に起源を求めてみよう。Doyle は，Swanson らがいかに探索者に心理的連想ネットワークを想起させるかという問題に関連して，シソーラスの概念を提示したのが，彼の意味地図の出発点であったと次のように述べている。

Swanson らは，同義語や関連語のシソーラスをこの問題の解決策として示した。連想地図は，ある意味で，この解法の拡張である。これは，巨大な自動的に導き出されたシソーラスである。このような地図を眼前にさし示されることによって，探索者は自分の思っているよりもずっと良い連想ネットワークを得ることができる。[34]

Doyle は，人間による索引と探索が，機械より優れた結果を生むという立場にたち，文章に書かれた連想結果も人間の連想行為の所産として，この逆関係も成りたつと考えた。したがって，彼の意味地図は，各索引語間の関連の強さや概念を図型的に表示して，索引作業を補助しようとするものである。

また，Artandi[35] が1961年に可能性を示した図書の巻末索引の自動作成についても大きな進歩はみられず，

やはり半自動，コンピュータ補助索引への移行がみられる。

1966年，IBM の Carney が示したシステム[36]では，コンピュータが重要語候補の選出，索引語とその語を含む文の印刷，本文内所在の明示，語の正順化，のちの検索用ファイルの保持などに利用された。このコンピュータ利用の結果，経済的コストも人間だけによるより低いとされた。[37] コストの点では，American Documentation の索引にコンピュータ補助索引を適用した，Hines と Harris[38]も，同じく経済的に好ましいとしている。

また Borko は SAINT (Semi-Automatic Indexing of Natural Text) システム[39]を発表して，図書巻末索引に相互介在の概念を導入した。相互介在とは，巻末索引の作成を連続した処理過程と考え，人間とコンピュータが相互に知的補助と事務的処理を分けながら，それぞれの中間結果を受けとりあうというものである。

最近の開発としては，ISI (Institute of Scientific Information) 発行の *Current Contents Weekly Snbject Index* 作成にコンピュータ補助索引が適用されている例[40]がある。

近年とくに注目すべき点は，コンピュータ技術一般におけるオンライン化と，それに伴う会話方式の普及が，コンピュータ補助索引の実用化をさらに促進していることである。この種の会話型システムとしては，IBM の Bennett が開発した，NSF (Negotiated Search Facility)[41]と，実用システム例としての SSIE (Smithsonian Science Information Exchange, Inc.) の索引方式が知られている。[42]

B. 内部索引法

自動索引の困難性を理解するには，前節でふれた図書巻末索引の特色を，自動索引の見地から見なおしておく必要がある。

一般に，索引はある文献コレクションを対象として，そこに含まれる個々の文献に対して行なわれるものであるが，このため，当該文献の外部（文献コレクション）を意識した索引という意味から，外部索引として，図書巻末索引と区別できる。これに対応して，後者は内部索引とよぶことができる。なぜなら，巻末索引は，その当該文献たる図書，すなわち，著者のある時点における一作品をほとんど閉鎖的に唯一の索引対象として考えられるからである。もちろん，読者の探索要求から生じる，語法的，意味論的差異の調整が必要であることはいうまでもないが，これも外部索引との比較で考えれば，明らかに容易に対処しうる。外部索引の場合，文献コレクションには，いく種類もの文献タイプが混在するかもしれないし，文献によって主題がばらつくために用語もまちまちであろう。文献コレクションが十分に限られた分野に集中している場合，専門的・技術的用語はより広い分野におけるよりも，ずっと明確な意味を持ってくる。カバーしなければならない語の数も急速に安定し，語彙の修正もずっと少なくてすむ。

内部索引は，こうした条件の極限的例といえよう。内部索引がもつ自然な有利性について，Maloney の指摘がある。

1．閉鎖的性格：システムの完成後，長い時間にわたって新しい語が入力される問題がない。ある図書の索引は，その図書が改訂されるまで，変更される必要がない。

2．限定的性格：索引そのものは量的にも索引深度上も過大でない。索引部分が 20 ページほどまでの図書なら，どんな図書でも全般にくまなく索引できるといえる。

3．個性的性格：文章表現上，観点や立場からくる妥協の必要がない。外部索引が直面するような，典拠リストやシソーラスなどへの慎重な配慮に比べて，内部索引には実質上，そのような努力が不必要であることが，実践における両者の相違を雄弁に物語っている。

4．特定的性格：索引の各項目の観点，バイアス，特定性は，その文章のものであり，より広いまたはより狭い，あるいはより神経質なまたはより高度な背景を持ってその資料を探そうとする探索者自身も，それを承知しているので，索引項目の選定や記入形式について影響を与えることがない。

5．自己検定的性格：索引によって要求する情報に到達できるかどうかの検定が，端的に可能である。

6．同義語問題からの解放的性格 (Synonym free)：索引中の語句は本文中の語句と同一である。[43]

このように，内部索引は単純な機械的処理技術が比較的容易に応用できる有利性を備えている。

上記の有利性は，同時に，索引の自動化の困難性を逆説的にいいあてている。そこで，Maloney の指摘をいいかえて，自動索引の困難性を整理すれば，次のようになろう。

1．開放的性格：外部索引（以下，単に索引という）は常に新しい概念の混入と，それに伴う，新しい語

句の発生に留意しなければならない。このような，索引対象の流動性から，辞書やシソーラス，個々の語句間の関係などは，常に開放的でなければならない。

2．拡散的性格：実際の索引現場の要請としては，情報の量的増大という意味あいが強いのは当然であり，複雑な自動分析(統計・構文分析など)を入力するすべての文献に対して行なおうとするには，なお無理がある。

3．汎用的性格：文献ごとの著者の，それぞれ固有の観点や立場から，微妙に，あるいは大きく，用語用法上の相違が生じる。したがって，これらを総括的に処理しようとする索引システムは汎用性をもたなければならず，そのために，あらゆる語彙統制上の煩雑さが加わる。さらに，コレクションの専門性が弱ければ弱いほど，語句の意味論的相違が生じ，また，これらを包含するシソーラスも肥大化する。

4．一般的性格：索引項目の選定などで，ある特定の文献だけに合致したような方式をとることはできない。探索者も，索引と実際の文献上の用語との差異を考慮しなければならない。

5．非検定的性格：通常，索引そのものと実際の文献とは，時間的・空間的にそれぞれ独立して存在している。このため，索引の利用者，つまり探索者は，索引によって示された情報の正確性，さらに，満足度をただちに検定できない。

6．同義語・同音意義語の問題：これは，3に含めて考えられるものだが，ある特定の文献で使用された語句は索引に使われた語句と必ずしも一致しない。

さらに，より本質的な問題点を指摘すれば，究極的には我々が我々自身のコミュニケーション活動について，未だ全く無知であるということである。統計的な解析にしても，もともと未解明な概念の連合に，統計的にどの程度近似できるかといった実験段階からは一歩も出ていない。また，変換文法でいわれるように，自然語が完全に数学的に記述可能であるという証明は，現在のところ皆無といってよい。また，自然語の構文構造が全て解明され，すべての構文単位と組合せを列挙できるという保証も全くない。

自動索引がこれまで対象としてきた，「書かれた純粋な自然語」は，やはりあまりにも複雑であった。

C. 実用化への逆説的アプローチ

ここまで，言語学者らによる自然語そのものの解明なしには，いかなる自動索引の研究も実用化は程遠い現状を認識したわけだが，では，自然語そのものが，前にあげたような複雑さをもたないとすれば，実用化は可能かもしれない。つまり，内部索引でみられたような有利性を，何らかの特殊な環境が性質的に備えている場合，自動索引が実用的システムとして成立可能なのではなかろうか。前に指摘したような困難性を克服していくのが，正統な研究努力とするならば，そのような困難性がはじめから少ない環境に，実用システムを作っていくことは，全くの逆説といえる。しかし，そうした逆説的アプローチは，もしそれが可能であるならば，自動索引システムの有効性を知らしめ，かつ，基礎研究の努力への刺激剤となりうるだろう。

そこで，ここでは，これまでの自動索引・コンピュータ補助索引が対象としてきた出版物形態ではないが，高度に専門化され，用語の意味論的問題も極めて少なく，また，使用される文体にも非常に限定的な特徴がみられる，病理学報告書の入力に，上記の特殊環境を求めてみよう。そこには，一般的には自然語と考えられないほどの特殊性をもつが，情報の発生者であり，かつ，利用者である病理学者間では，全く自然に使用されている言語環境がみられる。病院や医学研究機関では，近年とみにコンピュータが利用され，それに伴って，病理学報告書も効率的に機械検索したいという要求がたかまり，様々な研究がすすめられている。このなかで，報告書内容の入力を自動的に行なおうとする動きが目立つが，その作業は情報学的に索引作業とみなすことができる。

この病理学報告書の自動入力は，これまで病院内で独自に開発研究されてきたし，また，自動索引の研究者たちから注目されたこともなかった。しかし，上記のような逆説的アプローチが可能な，特殊な環境として，自動索引研究の文脈でとらえる価値があろう。

IV. 病理学報告書の自動入力処理

A. 病理学報告書とその自動入力

病院業務の自動化＝Hospital Automation を考えるとき，管理会計などの一般的な Business Applications と，特殊性の強い Medical Applications とに分けるのが好都合である。病理学報告書の処理は主として，後者の一部としての，病院における自然語処理という位置づけが可能である。

どのような分野でも，それまで日常的に行なわれてきた業務にコンピュータを導入する場合の，ひとつの大き

な心理的抵抗は，我々人間が普段使用している自然語から，ディジタルなコードへとコミュニケーション媒体を変える点にある。したがって，自然語データが主たるコミュニケーション手段になっているような適用業務では，これが大きな障害となる。

医学分野の専門家たちが，患者の治療のため，また仲間や学生に医学的概念を伝達するため，あるいは，研究成果を記述するために使用する医学データは，多くの場合，非数値的であり，ほとんど例外なく自然語の単語と数との組合せで作られており，しかも，通常名詞句かその変型で示される。病院内では，このような限定性があるとはいえ，自然語で書かれた文書類が数多く存在し，ファイル化されているが，従来の方法では検索要求に応じられないことが多い。こうした自然語情報，いいかえれば，医学記述データは，主としてラボラトリ・データとして発生し，報告作業という，一種の索引作業を経て，蓄積される。

病院や医系研究機関で自然語処理の研究が行なわれるようになったのは，ここ約10年間ほどの，ごく最近のことである。そして，興味深い事実は，この期間の文献調査を行なってみると，この種の研究はほとんど病理学報告書を対象としているものに集中していることに気づくのである。

さて，病理学者が通常扱う，記述的内容の報告書の主なものは，外科病理学報告書（Surgical pathology reports，図1参照）と剖検報告書（Autopsy reports）である。外科病理学報告書は，外科的に得られた組織標本の巨視的，微視的特徴の記述，および，病理学者の所見事項に対する解釈からなる。剖検報告書も同様であるが，これは複数の器官系の死後検査に関する詳細な記述である点が異なる。データ発生量の基準としては，ベッド数500の病院規模で1万～1万5千の外科病理学報告書，300～500 の剖検報告書が年間発生するとみられる。

機械を用いない，従来の報告の方式は，病理学者が組織検査時に所見事項を口述するというものである。この口述内容は報告書として書き写され，患者の病歴記録に加えられたり，他のファイルに蓄積される。これら報告書は患者の治療に直接，間接に役立つ，また医学研究教育に欠かせない，高い価値の情報を提供する。

病理学報告書に記入されるデータは，大きく分けて，デモ・データ（Demographic data＝人口統計学的データ）と医学診断ステートメントの2種類である。このうち，デモ・データは，身長，体重，年令などの数量的なものや，性別など択一的性格のデータがほとんどで，比較的単純な機械処理が十分可能である。これに対して，診断ステートメントは自然語でかかれたもので，コメント形式で補完的に書かれた部分を含むこともあるが，ほとんどが高度に専門化された病理学，ないしは医学用語で，それらが名詞句またはその変形として口述記入される。

伝統的には，これら診断ステートメントはICD（International Classification of Diseases）などの標準的コード表を用いて，コード化され，のちの検索要求に対応してきた。しかし，発生データ量の増大，コード化のために高度に専門的知識をもった専任者が必要であるなどの問題から，自動コーディングの可能性が研究されてきた。

コミュニケーション手段が，比較的言語学的困難性の少ない限定的自然語である点に着目して，かつ，上記の問題に対処しようと，はじめての試みを示したのは，Smith と Melton[44] であった。彼らの指摘によれば，形態学的診断と検索要求が安定的かつ単純であるため，自動化がたやすいのである。自動コード化システムの実現と各病院におけるファイルの整備から，疾病の地域的分布などを調査するためのコード化された保健データの国家的登録システムまで幅広い提案がなされたのが注目される。

B. 現存するシステム例

以下，病理学報告書の自動入力が現実にどのように行なわれているか，数例を検討する。

1. NCI 病理学情報システム

1966 年に Pratt と Thomas[45] が発表した，NIH（National Institutes of Health）の National Cancer Institute Information Processing System for Pathology Data（NCI 病理学情報システム）は，病理学報告書ファイルの生成・維持，探索，ファイル中のレコードの識別と検索，統計諸表の印刷などの機能をもっている。扱われるレコードは剖検，外科病理学，細胞病理学の各報告書で4万レコード以上あり，診断ステートメントは約16万にのぼった。このステートメントは SNOP（Systematized Nomenclature of Pathology）を用いた，病理診断の最少単位でとらえられている。SNOPは，米国病理学会が，病理学資料の組織化を目的として作成したもので，病理学的所見を，影響を受けた部位名（局所解剖学），疾病による組織の形態的変化（病理形態学），病原体や薬品など（病因学），および生理的・化学的異常または変化（機能）の4つのファセットで分析記述でき

UNIVERSITY OF ILLINOIS HOSPITAL
SURGICAL PATHOLOGY REPORT

NAME (LAST) (FIRST) (MIDDLE) | DATE 6 17 70 | FLOOR 7 E

UNIT NO.	SPEC	YR	A E Mo	A E Yr	R.	SEX	SERV.	SERVICE	DOCTOR
67 05 83	12 06745				W	M		SURGERY	CLARK

PATHOLOGIC DIAGNOSIS (0. 20 05)

1) Papillary transitional a enocarcinoma of urinary bladder deeply penetrating into the muscular layer.
2) All pelvic lymph nodes are negative for tumor metastasis.
3) Glandular hyperplasia of prostate.
4) Essentially normal vermiform appendix and jejunum.

Kodachrome	Interest	Prev. B	F. S	II CARD	DX SNOP Topog. 4 digits x=11 y=22	DX SNOP Morph. 4 digits	TUMOR CLASS. Neoplasm Nature	TUMOR CLASS. Hist. Origin	TUMOR CLASS. Behav.
				1					
				2					
				3					
				4					
				5					

PATHOLOGIC FINDINGS:

GROSS: Submitted fresh is an urinary bladder with an intact prostate, along with a segment of both the right and left ureters, right and left iliac lymph nodes, vermiform appendix, aortic lymph nodes and a segment of ileum. The serosa of the urinary bladder is smooth and glistening, except for a puckered area, along its posterior surface which measures in its greatest diameter, 2 cm. The entire urinary bladder measures 10x8x5 cm. An ill-defined mass can be palpated along its posterior wall. On opening the urinary bladder, a large, cauliflower-like tumor mass is seen arising from the posterior wall. This measures 7x5x3 cm. This has a grey-white, nodular appearance, with patchy areas of hemorrhage. The prostatic urethra is patent. The prostate measures in its greatest dimension, 4x 2 cm. Transection reveals its cut surface to have a milky secretion. The seminal vesicles are unremarkable. The right ureter measures 3 cm. in length, by 4 mm. in its diameter while the segment of left ureter measures 4 cm. with a greatest diameter of 5 mm. The right iliac lymph nodes vary in size from 3 to 4 mm. in greatest diameter. This is true of the left iliac lymph nodes. Transection of many of them do not reveal them to be involved by tumor tissue. The vermiform appendix measures 7 cm. in length with a greatest diameter of 7 mm. The serosal blood vessels are congested. Otherwise this is unremarkable. The thickness of ileum measures 6 cm. in length and in its circumference 5 cm. The mucosa has its usual pattern, and no gross evident lesions are noted. A frozen section which is done on one of the aortic lymph nodes, reveals it to be "negative for malignancy". The aortic lymph nodes vary in size from 3 to 4 mm. in diameter. None of them on transection appear to be involved by tumor tissue.

V. TEVES, M.D

MICRO: Sections along the tumor mass of the urinary bladder, show the tumor to be made up of broad masses of transitional cells with their nuclei displaying pleomorphism. The tumor has penetrated deep into the bladder wall and the tumor cells here have undergone keratinization simulating squamous cells. Sections of the ureters do not show any significant pathologic changes. The right and left pelvic lymph nodes do not show any evidence of metastasis. Sections of prostate show glandular hyperplasia. No tumor tissue present. Sections of the seminal vesicles show the lining epithelium to be hyperplastic. Sections of appendix and jejunum do not show any pathologic changes

J. TADANO, M D.

6 25 70

SIGNED C. A. KRACKOWER, M.D.
(FOR PATHOLOGY)

図 1 外科病理学報告書（イリノイ大学病院の例）

るという立場にたつ。例えば，“Metastatic choriocarcinoma in lungs and liver”なる所見は，

T2800 M8809 E0000 F0000 LUNGS, CHORIOCARCINOMA

T5600 M8809 E0000 F0000 LIVER, CHORIOCARCINOMA

のように，局所解剖学 (T) のファセットが2種 (lungs, liver) あるため，2つのエントリーに分けて記述される。

Graepel[46] らが指摘するように，このようなコーディング構造は，SNOPを使う者が自分たち自身の検索目的に合わせて設定するための道具にとどまるという意味から，ICDのような用語表とは異なる。また用語間の関係

THE MONTREAL GENERAL HOSPITAL

DEPARTMENT OF PATHOLOGY

Name: Martin, Mrs. F. Room No.: 506 Unit No.: 932347

Sex: F Age: 57 Service Dr.: Norman, K.C. Serial No: S.74-2300

Date Received: February 20/74

1)
CN: Diarrhea

2)
CD: Carcinoma of rectum

3)
OP: Anterior Resection

Specimen: Rectosigmoid.
The specimen is labelled "bowel, perineal part". The specimen consists of two segments of bowel, one appears to be the large bowel with the anus measuring 10.5 cm. in lenght, and 7.5 cm. in circumference. Three centimeters from the anal margin is seen an ulcerated circular lesion, 5 cm. in diameter with rolled edges and a grayish necrotic looking base. The cut surface of the ulcer reveals firm yellowish-white irregular area extending into the serosa. Seven nodular structures yellowish-white in color are identified. Probably lymph nodes measuring .4 cm. in diameter to 1 cm. x .4 cm. Representative sections are taken: Section No. 1 - from the proximal resection margin. Section No. 2 - from the distal resection margin. Sections 3 & 4 -from ulcer. Section 5 & 6 - from lymph nodes. The second segment of bowel measures 13 cm. in length, and 5 cm. in circumference. The external surface and the wall are unremarkable, although the mucosal surface reveals a raised structure, measuring .3 cm. in diameter. The rest of the mucosa is unremarkable. Representative sections are taken: Section No. 7 & 8 - from the resection margin. Section No. 9 from the raised lesion. Section No. 10 - from two nodular yellowish-white structures probably lymph nodes, firm and measuring 1 cm. x .3 cm. and 1 cm. x .5 cm.

DIAGNOSIS: Anterior Resection: Infiltrating adenocarcinoma of rectum extending through muscularis.
Lymph nodes and resection margins free of tumor.

Note: Very vigorous immune reaction apparent in local lymph nodes dissected.

注 1) CN—臨床注記
2) CD—臨床診断
3) OP—手術

図 2 モントリオール総合病院における従来の病理学報告書の例

を表現していないので，この探索構造の設定のための配慮を加えることが，SNOP を病理学情報システムにおけるシソーラスとしての位置づけに必要不可欠である。

このように，主として名詞句で示される病理学的所見の SNOP コードへの自動コード化を全面的にめざしている点が，このシステムの注目すべき点である。

2．モントリオール総合病院

カナダのモントリオール総合病院では，隣接のマッギル大学付属病院施設を包含する前提でコンピュータによる病理学報告書処理のため，CISP (Computerized Information System for Pathology) を設計した。[47] CISP の対象は，NCI 病理学報告書システムと同様で，このうち最も多い外科病理学報告書は年間約12万件にのぼる。

この外科病理学報告書は，識別，臨床，概略記述，病理診断の4セクションから構成され，臨床および病理診断の両セクションについては，システムに入力された文章が分析され，内部辞書を用いて，統制語彙にコード化される。この辞書は，単純ではあるが，階層関係を示せるように構成されている点，同義語やサブカテゴリーなどを統制できるように配慮されている点など，シソーラス機能を備えているうえ，その作成にあたり，ICDA (ICD Adaptations＝ICD 米国適用版) と SNOP を組合せてあることが注目される。

SNOP は，この分野の最も普及した用語集であるが，各々の診断ステートメントを表現しなおさなければならないこと，記入項目が過度に詳細化してしまうこと，臨床と病理との相関で必要な臨床医学用語が欠如していることなど，不都合な点があるので，これら欠点の一部を ICDA によって臨床診断面のコーディングの容易化をはかっている。

3. UCLA 医学部付属病院

病院における自然語処理システムとして，最大規模かつ，最も広い対象分野をもつのが，UCLA（カリフォルニア大学ロスアンゼルス校）医学部付属病院の Natural Language Retrieval System である。[48] このシステムはやはり病理学報告書処理を出発点としており，現在解剖・骨髄・神経放射線・核医学を含めた5分野の報告書類を対象としている。

前出の Smith と Melton の指摘をうらづけるように，次の諸点が当システム開発の理由となっている。

```
                THE MONTREAL GENERAL HOSPITAL
                  DEPARTMENT OF PATHOLOGY

SURGICAL NO.: S.74- 2300    UNIT NO.:  932347    ROOM NO.: 506
NAME: MARTIN,MRS. F                    SEX: FEMALE     AGE:  57
SERVICE DR.: NORMAN K C
SPECIMEN DATE: FEBRUARY 20/74
DN:  1. DIARRHEAL DISEASE — DIARRHEA
CD:  1. CARCINOMA(RECTUM)
OP:  1. COLECTOMY AND ENTEROSTOMY -- RESECTION OF COLON , SEGMENTAL -- A
        NTERIOR RESECTION
SP:  SPECIMEN: RECTOSIGMOID
          THE SPECIMEN IS LABELLED "BOWEL,PERINEAL PART". THE SPECIMEN CONSISTS
     OF TWO SEGMENTS OF BOWEL, ONE APPEARS TO BE THE LARGE BOWEL WITH THE ANUS
     MEASURING 10.5 CM. IN LENGTH , AND 7.5 CM.  IN CIRCUMFERENCE . THREE CENTIMETERS
     FROM THE ANAL MARGIN IS SEEN AN ULCERATED CIRCULAR LESION, 5 CM. IN DIAMETER
     WITH ROLLED EDGES AND A GRAYISH NECROTIC LOOKING BASE. THE CUT SURFACE OF THE
     ULCER REVEALS FIRM YELLOWISH-WHITE IRREGULAR AREA EXTENDING INTO THE SEROSA.
     SEVEN NODULAR STRUCTURES YELLOWISH-WHITE IN COLON ARE IDENTIFIED , PROBABLY
     LYMPH NODES MEASURING .4 CM. IN DIAMETER TO 1 CM. X  .4 CM. REPRESENTATIVE
     SECTIONS ARE TAKEN: SECTION NO.1 - FROM THE PROXIMAL RESECTION  MARGIN .
     SECTION NO.2 - FROM THE DISTAL RESECTION MARGIN.
     SECTIONS NO. 3&4 -  FROM ULCER .
     SECTION NO.5&6 - FROM THE LYMPH NODES. THE SECOND SEGMENT OF BOWEL MEASURES 13
     CM. IN LENGTH , AND 5 CM. IN CIRCUMFERENCE . THE EXTERNAL SURFACE  AND THE WALL
      ARE UNREMARKABLE , ALTHOUGH THE MUCOSAL SURFACE REVEALS A RAISED STRUCTURE,
     MEASURING  .3 CM. IN DIAMETER. THE REST OF THE MUCOSA IS UNREMARKABLE .
     REPRESENTATIVE SECTIONS ARE TAKEN: SECTION NO.7&8 - FROM THE RESECTION MARGIN.
     SECTION NO.9 FROM THE RAISED LESION.  SECTION NO.10 - FROM TWO NODULAR YELLOWISH
     -WHITES STRUCTURES PROBABLY LYMPH NODES, FIRM AND MEASURING 1 CM. X .3 CM.  AND
     1 CM. X .5 CM.
PD:  1. RECTUM , NOS
          CARCINOMA -- ADENOCARCINOMA
            INFILTRATING
            EXTENSION TO MUSCULARIS
            RESECTION MARGINS FREE OF TUMOR
            NO LYMPH NODE INVOLVEMENT
NT: VERY VIGOROUS IMMUNE REACTION APPARENT IN LOCAL LYMPH NODE DISSECTED
```

図 3　モントリオール総合病院における病理学報告書のコンピュータ出力例

1．適合する文章の検索のためには，特定のタームまたはその集合が存在するかどうかだけで充分文章を分別することができ，綿密な構文論的・意味論的分析を行なう必要がない。

2．文章には，ほとんど必ず，それ自体の意味が明瞭なディスクリプタがある。

3．オリジナルの英語そのままをコンピュータ内に蓄積して，検索時に全文探索するのは時間的・経済的に不利である。

4．これに対して，コンピュータによる自動コーディングは，自然語のまま入力でき，また自然語に変換し出力できる。

このシステムの特徴は，シソーラスが辞書とは独立して備えられている点である。Lamson-IBM シソーラスと通称されるこのシソーラスは，2進整数で表わされたキーワード内部コード間の関係を統制し，辞書はそれら内部コードと自然語のインターフェイスとして用いられている。これら関係概念には，

1．類義語クラスを設けて語をグループ化する類義性

2．類義語クラス同志を結びつけることで親セットと子セットの間の従属性

3．論理関係とよばれる相互結合ネットワーク

などが含まれる。ディスクリプタの更新は，科学用語の流動性を考慮して，容易に行なえるよう設計されている。1974年7月現在，UCLA 病理学シソーラスと UCLA 核医学シソーラスの2種類が用意されている。

C. 病理学報告書システムの問題点と今後

これまでみてきたように，一般の文献に比べて言語学的に困難性が少ないことから，情報処理技術上の問題点も少ないと考えがちであるが，索引精度の要求については，一般の医学文献情報よりはるかに高く，診断情報の索引性能の信頼性は，索引上のノイズの容認度で表わしたとき，1％を充分に下回わる必要があるという意見[49]がだされている。

診断の確信度や否定的診断，すなわち，〝ある病理診断の可能性あり〟や〝病理診断なし〟についても，検索精度の要求によって，扱いをはっきりさせておく必要がある。一般には，キーワードの存在の有無だけで，単純に肯定的病理診断を索引し，蓄積するにとどめるが，より高度なシステムでは，これらの問題を配慮する必要がある。SNOP の場合は，コードを別に割りふる方法をとっているが，余裕コードが現在でも不足している SNOP にとって，さらに柔軟性を低くしているともいえよう。自動入力では，否定的診断を蓄積するか否か以前に，否定的診断であるかどうかの認識が問題となる。否定語とキーワードの共出現だけでとらえるアルゴリズムでは，重要なキーワードの索引もれをひき起こす場合もあり，この点，入力のための何らかの口述規制（いわゆる疑似自然語入力）が好結果をもたらす。

病理学関連の用語集はすでに，SNOP, Lamson-IBM シソーラス，ICDA など数多く存在し，こうした多様な選択性は，辞書やシソーラスの自己開発をも含め，システム設計上の混乱を招くばかりでなく，将来的な地域医療システムや連邦規模の病理学情報交換への統合を考えたとき，その適合性 (System compatibility) の点で大きな障害となる。しかし，一般的傾向として，語彙標準化のための努力は，徴候名の分野で，ドイツの DOFONOS (German Syndrome Identification and Information System)[50] の活動がみとめられる程度で，米国では病理学だけでなく，一般に活発でない。

自動コーディングの際は，決定論的な語の割当が適するため，統計的手法が適用された例は少ないが，登録すべきキーワードやストップ・ワードの選定にあたっては，統計的手法が利用できよう。ただし，キーワードの選定は，病理学専門家の判断なしには実用上行ない得ないので，ストップ・ワード選定に関して，ノイズの低減化，索引もれの回避などのための有効な判定資料として，例えば Harter の β 測度が利用できよう。初期の自動索引研究で盛んに用いられた相関分析は，全く病理学研究領域そのものに属するものであり，情報システムが直接関与すべき性質のものではない。

経済的観点からは，本来，情報検索システムにおいて，その大部分のコストを占める入力のための人件費を節減する意味で問題ないのだが，このような効果は，大きなシステムではじめて得られるのが現状である。この点では，病理学報告書の処理要求が，大規模な病院や研究施設に集中することから，現在の方向が必ずしも好ましからざるものとはいえないだろう。

さらに，自動入力においては，SMART システムのような複雑な句構造解析の準備は，かえってシステムの経済的実用効率を低下させるおそれがあるので，構文分析的には，分詞句，前置詞句の認定や句構造の正当性検定等にとどめるのが実際的といえようし，また診断情報の言語的性質からも，この程度の解析が適しているだろう。

今後の報告書処理は，病院自動化の一般的すう勢に沿

って,オンライン化がすすむと考えられる。したがって,自動入力も会話型で行なわれる,コンピュータ補助索引タイプを指向しよう。この場合,辞書やシソーラスをオンラインで照会し,人間が画面表示をとおしてモニタするといった光景が現実化してこよう。

V. 結　　論

これまで,自動索引研究の足取りをたどり,それを踏まえて,病理学報告書の自動入力を観察したわけだが,このような病院における制限された情報活動が,文献の自動索引に対してどのような具体的貢献をもたらすかを考察し,本論文の結論としたい。

それは,次の3点に集約されると考える。

1．自動索引法の基礎的研究成果への貢献
2．自動索引の実用化促進
3．図書館・情報学の一分野としての自動索引研究の他領域への直接的貢献

1の例としては,否定の扱いがある。構文構造が複雑すぎ,語の意味論的あいまい性が高い場合,否定語の扱いに様々の構造解析と例外の設定が必要とされ,また,その結果の信頼性もはなはだ低かったのに対して,逆説的アプローチによって,限定された分野では結果の評価も容易であり,少なくとも限られた句構造パターンでの否定の扱いができるようになった。このように,結果が評価しやすいことは,諸手法の有効性検定の材料提供に好都合である。

本論文で最も強調したい点は,自動索引の実用化の問題である。実現不可能であるとして研究そのものが停滞気味であった自動索引は,このような分野で着実に育ちつつある。その高い実用性が認められるようになってこそ,また基礎研究の活気も生まれようし,なんといっても社会的に自動索引を受けいれる素地を作り,確立していくことが肝要である。その意味でも,病理学報告書や他の病院内部資料の処理に限らず,積極的に実用化可能な環境の開拓をすすめるべきである。同時に,文献資料の処理という,自動索引の従来的なとらえ方からすすんで,あるいは,図書館という一情報システム類型にとらわれずに,他の情報システム,他の領域を積極的に包含していくことによって,自動索引を広く社会的に知らしめ,その有効性,汎用性を示す必要もある。

言語の解明は,我々人類が,我々自身の文明を知りつくせぬと同様に,永遠に果たしえない課題かもしれない。しかし,深遠な言語文明の謎解きと同時に,我々の目前には,情報の絶対量の過剰増加という問題のあることを忘れるわけにはいかない。実用的な自動索引システムの確立は,なんとしても現在の時点で必要なのである。

最後に,本論文では,文脈を米国に限定したが,日本語の自動索引を考えれば,漢字の機械処理等,近年めざましい発達はあるものの,入力における絶対的な煩雑さ,高コストは否定しがたく,また自動索引研究例も少ないなどの点から,米国におけると同様の論議は避けるべきで,近い将来,日本独自の問題提起がなされるのを待ち,今後の調査研究の重要性をくりかえし強調したい。

1) Meadow, C. T. *The analysis of information systems.* 2nd ed. Melville, Los Angeles, 1973. p. 15.
2) Baxendale, P. B. "'Autoindexing' and indexing by automatic processes," *Special libraries*, vol. 56, no. 10, Dec. 1965, p. 715.
3) Salton, G. ed. *The SMART retrieval system; experiments in automatic document processing.* Prentice-Hall, Englewood Cliffs, New Jersey, 1971. p. 545.
4) Borko, H. The construction of an empirically based mathematically derived classification system 〈*American Federation of Information Prcessing Societies, proceedings.* Spring Joint Conference. 1961〉 p. 279-81.
5) Luhn の研究については, Schulz, C. K. ed. *H. P. Luhn; pioneer of information science; selected works.* Spartan Books, New York; MacMillan, London, 1968. 320 p. を参照するのが便利である.
6) Swanson, D. R. *An experiment in automatic text searching, word correction and indexing. Phase 1. Final report.* Thompson-Ramo-Wooldridge, Inc. Canoya-Park, Calif., 1960. p. 36+5 (Report C82-OU4)
7) Maron, M. E. "Automatic indexing; an experimental inquiry," *Journal of the association for computing machinery*, vol. 8, 1961, p. 401-17.
8) Borko, *Loc. cit.*,
9) Borko, H. and Bernick, H. "Automatic document classification," *Journal of the association for computing machinery*, vol. 10, 1963, p. 151-62.
10) Stiles, H. F. "The association factor in information retrieval," *Journal of the association for computing machinery*, vol. 8, 1961, p. 271-9.
11) Baker, F. B. "Information retrieval based

upon latent class analysis," *Journal of the association for computing machinery*, vol. 9, 1962, p. 512-21.

12) Williams, J. H. A discriminant method for automatically classifying document 〈*American Federation of Information Processing Societies, proceedings*. Fall Joint Computer Conference, vol. 24, 1963〉 p. 161-6.

13) Damerou, F. L. "An experiment in automatic indexing," *American documentation*, vol. 16, no. 4, Oct. 1965, p. 283-9.

14) Bookstein, B. and Swanson, D. R. "Probabilistic models for automatic indexing," *Journal of the American Society for Information Science*, vol. 25, no. 5, Sept.-Oct. 1974, p. 312-8.

15) Harter, S. P. "A probabilistic approach to automatic keyword indexing. Part I. On the distribution of specialty words in a technical literature," *Journal of the American Society for Information Science*, vol. 26, no. 4, July-Aug. 1975, p. 197-206.

16) Harter, S. P. "A probabilistic approach to automatic keyword indexing. Part II. An algorithm for probabilistic indexing," *Journal of the American Society for Information Science*, vol. 25, no. 5, Sept.-Oct. 1975, p. 285-9.

17) Artandi, S. Automatic indexing of drug information 〈*American Documentation Institute, proceedings*. 1967 Annual Meeting, Oct. 22-27, 1967, New York, N. Y., 1967〉 p. 148-151.

18) Artandi, S. *Automatic indexing of drug information. Project MEDICO. Final report*. Graduate School of Library Service, Rutgers University. New Brunswick, New Jersey, 1975. vii, 24 p. (National Library of Medicine Grant LM-94)

19) Artandi, S. and Wolf, E. F. "The effectiveness of automatically generated weight and and links in mechanical indexing," *American documentation*, vol. 20, no, 3, July 1967, p. 200.

20) *Ibid*.

21) Earl, L. L. "Experiments in automatic extracting and indexing," *Information storage and retrieval*, vol. 6, 1970, p. 313-34.

22) Earl, L. L. *Automatic informative abstracting. Part I. Experiments in the use of syntactic information in automatic extracting and indexing. Final report*. Lockheed Missiles and Spaces Co., Inc. Palo Alto Research Laboratory, Palo Alto, Calif., 1973. 199 p. (INIS AD-762456; Report LMSC-D350104; Contract N000-14-70-C-0239)

23) *Ibid*., p. *A*-49.

24) 統治語とは，その語の存在が文章内で何らかの特徴的構造を示す性質のある語のこと．例えば，統治語believeは，以下に続く文章構造を1. 名詞，2. that＋節，3. what＋節，4. in 名詞，5. in what＋節，6. 名詞 to be 名詞のように規定する．

25) Hillman, D. D. "Customerized user services via interactions with LEADERMART," *Information storage and retrieval*, vol. 9, 1973, p. 583-96.

26) Gardin, J. C. *SYNTOL*. Graduate School of Library Service, Rutgers University. New Brunswick, New Jersey, 1965, 106 p.

27) Salton, G. *Automatic information organization and retrieval*. McGraw Hill, New York, 1968. 514 p.

28) Salton, G. *The SMART retrieval system; experiments in automatic document processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1971. 556 p.

29) Bernier, C. L. "Indexing and thesauri," *Special libraries*, vol. 59, no. 2, Feb. 1968, p. 102.

30) Fangmeyer, H. *Semi-automatic indexing—state-of-the-art*. AGARDograph report, AGARD, Paris, 1974. p. 4.

31) Klingbiel, P. H. "Machine-aided indexing of technical literature," *Information storage and retrieval*, vol, 6, 1970, p. 29 の抄録文.

32) Scheffler, F. L. and Smith, R. B. *Document retrieval system operations including the use of microfiche and the formulation of a computer aided indexing concept. Final summary report*. Dec. 1, 1967 to Nov. 30, 1968, Dayton University Reserch Institute, Dayton, Ohio, 1968. 50 p. (CFSTI AD-686 804)

33) Doyle, L. B. "Semantic road maps for literature searchers," *Journal of the association for computing machinery*, vol. 8, 1961, p. 553-8.

34) Doyle, L. B. "Indexing and abstracting by association," *American documentation*, vol. 13, 1962, p. 378-90.

35) Artandi, S. *Book indexing by computer*. Graduate School of Library Service, Rutgers University. New Branswick, New Jersey, 1963. (Ph. D. Dissertation) 206 l.

36) Carney, F. J. Computer assisted index preparation 〈*American Documentation Institute, proceedings*, 1966 Annual Meeting. Oct. 3-7, 1966. Santa Monica, Calif., 1966〉 p. 329-38.

37) *Ibid*., p. 331.

38) Hines, T. C., Harris, J. I. and Colverd, M. "Experimentation with computer assisted indexing: *American Documentation* vol. 20," *Journal of the American Society for Information Science*, vol. 21, no. 6, Nov.—Dec. 1970, p. 402-5.

39) Borko, H. "Experiments in book indexing by computer," *Information storage and retrieval*, vol. 6, no. 1, May 1970, p. 5-16.
40) Neufeld, M. L. *et. al.* "Machine aided title word indexing for a weekly current awareness publication," *Information storage and retrieval*, vol. 10, no. 11-12, Nov.—Dec. 1974, p. 403-10.
41) Bennett, J. L. "On-line access to information; NSF as an aid to the indexer/cataloger," *American documentation*, vol. 20, no. 3, July 1969, p. 213-20.
42) Hersey, D. F. *et. al.* On-line retrieval and machine-aided indexing in a large data base of ongoing research information 〈Waldron, H. J. and Long, F. R. *eds. Proceedings of ASIS*, vol. 10. Annual Meeting, Oct. 21-25, 1973. Los Angeles. Greenwood Press, Westport, Conn., 1973〉 p. 89-90.
43) Maloney, C. J., *op. cit.*, p. 83.
44) Smith, J. C. and Melton, J. S. "Automated retrieval of autopsy diagnoses by computer technique," *Methods of information in medicine*, vol. 2, no. 5, 1963, p. 85-90.
45) Pratt, A. W. and Thomas, L. B. "An information processing system for gathology data," *Pathology annual*, 1, 1966, p. 1-21.
46) Graepel, R. H., Henson, D. E. and Pratt, A. W. "Comments on the use of the systematized nomenclature of pathology," *Methods of information in medicine*, vol. 14, no. 2, 1975, p. 72-5.
47) Hercz, L., Laszlo, Ch. A. and Reesal, M. R. "A computerized information system for pathology," *Methods of information in medicine*, vol. 14, no. 4, 1975, p. 182.
48) Okubo, R. S. *et. al.* "Natural language storage and retrieval of medical diagnostic information. Experience at the UCLA Hospitals and Clinics over a 10-year period," *Computer programs in biomedicine*, vol. 75, 1975, p. 105-30.
49) Long, J. M., Barnhard, H. J. and Levy, G. C. "Dictionary building and stability of a word frequency in a specialized medical area," *American documentation*, vol. 18, no. 1, 1967, p. 21-5.
50) Leiber, B. "The German syndrome identification and information system (DO FoNos)," *Methods of information in medicine*, vol. 14, no. 2, 1975, p. 67-72.

参 考 文 献

Collen, M. F. ed. *Hospital computer systems*. Wiley, N. Y. 1974, 768 p.
Pratt, A. W. "Medicine, computers and linguistics," *Advances in biomedical engineering*, vol. 3, 1973, p. 97-140.
Sparck Jones, K. "Progress in documentation; automatic indexing," *Journal of documentation*, vol. 30, no. 4. Dec. 1974, p. 393-432.
Stevens, M. E. *Automatic indexing; a state-of-the-art report* (with list of references cited and selected bibliography) U. S. National Bureau of Standards, Washington, D. C., 1965. 220 p.