Library and Information Science No. 31 1993

複数領域における日本語原著論文の機能構造分析: 構成要素カテゴリの自動付与

Functional Structure Analysis of Research Articles Selected from Three Specialties: Automatic Category Assignment

神　門　典　子
Noriko Kando

Résumé

This paper conducts functional structure analysis of research articles, using the Categories, the author previously proposed as a framework for analysing informational content of research articles. Automatic assignment of the Categories is conducted to the corpus of research articles selected from three specialties, using matching of the lexcal-clue-combination-patterns and stochastic rules. The result is, 86%-91% of Category fixed sentences are assinged correct Categories automatically. And through analysis of errors and related studies followings are suggested to improve efficiency of automatic assignment; (1) to consider inter-sentences relationship, (2) to adopt syntactic patterns of lexical clues combination, (3) to introduce the wholistic processing strategy, (4) to examine on modification of the Categories, which are needed content evaluation.

神門典子: 慶應義塾大学大学院博士課程, 東京都港区三田 2-15-45 (日本学術振興会特別研究員).
Noriko Kando: Graduate School of Library and Information Science, Keio University, 2-15-45, Mita, Minato-ku, Tokyo 108, Japan.
(*Research Fellow of the Japan Society for the Promotion of Science*)
1994年2月9日受付

I. 構成要素カテゴリとは

構成要素カテゴリとは，情報メディアの伝達内容の特性を，領域に関わらず，メディアの種類ごとに共通にとらえる基盤として筆者が提案したものである[1]。各カテゴリは，情報メディア内の部分が果たしている役割や機能を表わす。

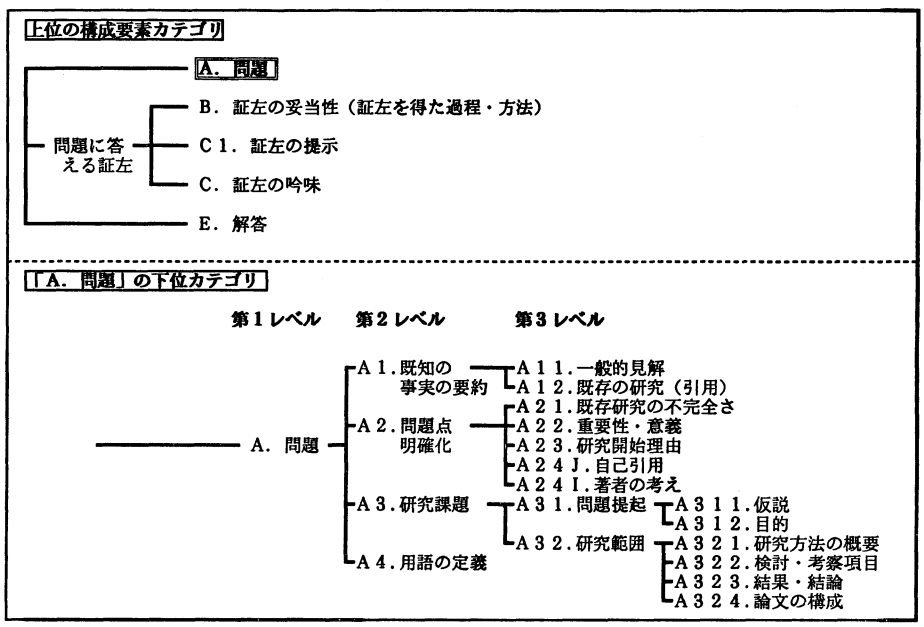
情報メディアの一種である原著論文を分析するために，各種の論文執筆マニュアルと複数領域の実際の原著論文の内容の分析を通じて81個のカテゴリからなる一連の階層構造を持った構成要素カテゴリを設定した[1]。その中で，上位のカテゴリと本稿で対象とする「A. 問題」の部分を第1図に示した。

論文内でのカテゴリの有無や出現順序によって，機能構造，すなわち原著論文が伝達している内容を機能という側面からとらえた構造を記述できる。カテゴリの出現型は，基本的な型といくつかの部分的な型の組合せとしてとらえ，部分的な型の繰り返しやその一部の脱落を許容して大局的にとらえると，いくつかの特徴的な型に類型化することができた[1]。

カテゴリを付与する単位は原則として文であり，全ての文にカテゴリを付与する。必要に応じて1文内の部分にも付与できる。局所的な役割を表わすカテゴリと全体から見た役割を表わすカテゴリとを二重に付与する「入れ子構造」も認めている。

構成要素カテゴリの概要と分析基準の再検討については，別稿にて報告している[2]。

既報[3]にて概観したように，図書館・情報学では，従来から，情報メディアの特性を明らかにするために，あるいは抄録作成，索引作成，情報検索などへ応用することを想定して，情報メディアの伝達内容の構造が分析されているが，分析方法や分析の詳しさは研究の目的によって様々である。また，学術論文の特性を明らかにするために，言語学，科学社会学，科学史でも，論文の機能構造が手作業で分析され，論文の機能構造に関する知識があると論文内容の記憶が促進されるという報告もある[3]。日本語で書かれた論文を対象とした自然言語処理研究において全文データベースの検索[4]やブラウジング[5,6]へ応用する目的で，機能構造の自動分析が試みられている。



第1図　原著論文を分析するための構成要素カテゴリ：最上位のカテゴリと「A. 問題」の下位カテゴリ

それに対し，構成要素カテゴリを用いた分析は，特定の応用のためのものではなく，情報メディアの伝達内容の特性をとらえることを第一の目的としているが，全文データベースへの応用をはじめとして，情報メディアの蓄積，検索，加工に関してさまざまな応用も可能であると想定している[1]。

しかしながら，構成要素カテゴリの付与は時間がかかる作業なので，実用化するには付与の自動化が必要である。日本語のC型肝炎論文を対象として，特徴的な手がかり語の出現確率とカテゴリの出現パターンに基づくカテゴリの遷移確率を用いて「A. 問題」の下位カテゴリを対象として予備的に自動付与を試みたところ，95.2%の精度で成功し，その実現可能性が示された[7]。一方，領域によって，論文の構造の複雑さや記述の仕方の特徴が異なることが報告されている[1)8)9)]。

そこで，本稿では，複数領域の原著論文を対象としたカテゴリ自動付与を試み，その問題点を明らかにし，言語表現という側面からの検討[2]を踏まえて，解決策を検討した。

II. 分析対象

分析対象は第1表に示したように，3つの領域から選択した日本語で書かれた原著論文である。いずれも別稿[2]で分析基準を再検討する際に用いたものであり，すでに全文に人手でカテゴリが付与されている。

第1表の中で，C型肝炎論文は前報[7]でカテゴリ自動付与の対象とした文献群である。本報は複数領域の論文を対象とした自動付与の問題点を検討することを目的としている。そこで，情報検索研究と対人認知領域の論文を分析対象に追加した。いずれの論文も全文を機械可読形態に変換して用いた。

また，著者名や抄録は除外し，論文の標題と本文だけを用いた。標題，論文中の章節の見出し，段落の開始，引用文献や図表の参照には，SGML (Standard Generalized Markup Language) に準拠したタグを手作業で付与した。カテゴリは，章節の見出しに関わらず，文自体の内容に基づいて付与し，標題と章節の見出し以外の全ての文に付与する。

前報[7]と同様に，論文の中心的課題を示すまでの部分に相当する「A. 問題」の下位にある諸カテゴリを対象とした。ここはさまざまな応用が考えられる部分である。たとえば，「A3. 研究課題」カテゴリが付与される部分は，いわゆる主題文に相当し，実際の索引作成者による索引作成過程のプロトコルでも重視されていた[10]ことから索引作成や自動索引への応用も考えられる。また，この部分から領域の現状把握に有用な文の抽出[7]も実際に試みている。

第1表　分析対象論文

	C型肝炎	情報検索	対人認知
論文数	50件	49件	38件
出版年	1991	1979-91	1985-91
長さ 全体	3,682文*	9,477文	6,723文
(1件あたり)	(73.6)	(193.4)	(176.9)
A問題ｶﾃｺﾞﾘ	293文	914文	1,482文
(1件あたり)	(5.9)	(18.7)	(39.0)

*: 句点「。」で区切られたものを1文とする。ただし，箇条書き部分では，句点が省略されているものも文とみなした。本文中の章節の見出しは除く。

第2表　構成要素カテゴリ別にみた，のべ付与カテゴリ数* とカテゴリ出現型

	C型肝炎	情報検索	対人認知
A11 一般見解	69	119	98
A12 既存研究	112	138	471
A21 不完全さ	34	141	68
A22 重要性	8	99	5
A24 著者立場	36	193	566
A311 仮説	−	−	74
A312 目的	8	32	94
A321研究概要	55	150	59
A322検討項目	14	13	31
A323結果結論	7	15	−
A324論文構成	−	43	−
A4 用語定義	−	15	28
論文ごとに見た構成要素カテゴリの出現型**(単位:論文)			
A1−A3	13	−	−
A1−A2−A3	28	2	−
(A1−A2) −A3	6	22	6
(A1-A2) - (A2-A3)	−	6	4
(A1−A2−A3)	−	9	22
(A2−A3)	3	9	−
A3- (A1-A2-A3)	−	1	6

* 一文の前半と後半でカテゴリが異なる場合は、前半部のカテゴリと後半部のカテゴリの双方に計数されているので、のべ付与カテゴリ数の合計は表1にしめした文の数より多い。

** 第二レベルのカテゴリにまとめている。論文内でカテゴリが出現した順に左から記述してある。(　) は () 内の部分が繰り返し出現する型を表わす。ただし、() 内の一部カテゴリの脱落を許容する。

領域別に特徴を見ると，第1表のごとく，情報検索と対人認知領域の論文は「A. 問題」カテゴリ部分が長かった。すなわち，C型肝炎では1論文あたり平均5.9文であったのに対し，情報検索は約3倍の18.7文，対人認知では約7倍の約39.0文であった。

また，カテゴリの出現も，第2表のごとく「A311. 仮

説」「A324. 論文の構成」「A22. 重要性・意義」などは領域によって差が見られた。カテゴリの出現型に関しても，第2表に示したように，C型肝炎論文では，「A1. 既知の事実」カテゴリのあとに「A3. 研究課題」カテゴリが出現する，もしくは「A1. 既知の事実」カテゴリのあとに，「A2. 問題点の明確化」カテゴリと「A3. 研究課題」カテゴリが出現するという単純な型がほとんどであった。それに対し，情報検索や対人認知領域は部分的なカテゴリの出現型の繰り返しが多く，特に対人認知領域では「A3. 研究課題」カテゴリが繰り返し出現する複雑な型の論文が多かった。

このように対象部分の長さ，カテゴリの有無や出現型が様々である複数領域の論文を対象とすることにより，分析に関わる問題点がより明らかになり，その問題点を解決することによって，自動付与の適応範囲を拡張することができると考える。

前報の枠組み		本報の枠組み
処理対象：C型肝炎　50論文（293文）		処理対象：3領域　計137論文（2，689文）
●確率的なルール		●確定的ルール＋確率的ルール
基準文献群（医学論文86件）中の①手がかり表現の出現確率	→	・条件の組合せによって確定する ・前文との連接関係を重視
②カテゴリの遷移確率（文単位・最詳細レベル）	→	・出現型として捉えたカテゴリ遷移（第2レベル）
③表層的パターン照合	→	●表層的パターン照合
④カテゴリ付与：文単位	→	●カテゴリ付与：文単位

第2図　自動付与処理の枠組み：前報と本報との比較

```
●もし、この文に「本稿」「本研究」「本報」のいずれかが出現するならば
      「A3.研究範囲」の下位カテゴリの得点を ＋0.38する。
      「A4.用語の定義」の得点を　＋0.21する。
      さらに、「そこで」が出現するならば、
          「A3.研究範囲」の下位カテゴリの得点を ＋0.41する。
      さらに、「定義」が出現するならば、
          「A4.用語の定義」の得点を　＋0.52する。
      さらに、「目的」が出現するならば、
          「A312.研究範囲」の得点を ＋0.67する。
                    ・
                    ・
                    ・
●もし、この文に「われわれ」「著者(ら)」「筆者(ら)」が出現するならば
      「A3.研究範囲」の下位カテゴリの得点を　＋0.3する。
      「A24.著者の立場・従来からの関心」の得点を　＋0.21する。
                    ・
                    ・

●もし、この文に「～ている。」が出現するならば、
      「A1.既存の事実の要約」の下位カテゴリの得点を ＋0.50する。
      「A21.既存研究の不完全さ」の得点を ＋0.09する。
      「A24J.自己引用」の得点を　＋0.30する。
                    ・
                    ・
```

第3A図　確率的ルールの例

III. 自動付与の枠組み

A. 前報で用いた自動付与の枠組み

前報では，以下の枠組みでカテゴリの自動付与を行なった（第2図参照)。まず，医学全領域から無作為に抽出した論文86件を基準文献群とし，それら文献に人手でカテゴリを付与し，その中での①手がかり表現の出現確率と②カテゴリの遷移確率とを求め，それに基づいて確率を用いた自動付与のルールを作成した。

手がかり表現の出現確率は次式によって算出した。基準文献群において出現数が少ない手がかり表現は，出現の仕方や機能が類似している語からなる小グループを作成し，そのグループごとに調べた。

$$\text{各手がかり表現（グループ）の出現確率} = \frac{\text{基準文献群中の当該カテゴリでの出現回数}}{\text{基準文献群全体での出現回数}}$$

カテゴリ遷移確率は，最も詳細なレベルのカテゴリに着目し，次式によって次文のカテゴリの確率を求めた。

$$\text{Xカテゴリから Yカテゴリへの遷移確率} = \frac{\text{基準文献群でXカテゴリの次にあるYカテゴリ文の数}}{\text{基準文献群でのXカテゴリ文の数}}$$

これらの確率を用いて自動付与のルールを作成し，基

●もし、この文に

「そこで」 ＋ ［本稿では／本論文では／本稿では／本報では／本報告では／本研究では／・／・／・］ ＋ ［測定した。／測定を行なった。／測定をした。／分析した。／分析を行なった。／分析をした。／分析を実施した。／・／・］

が出現するならば、この文は「A321.研究方法の概要」である。

●もし、この文に

「人名<r>は」 ＋ ［見いだしている。／発見している。／発見をしている。／明らかにしている。／確認している。／・／・］

または、

［最近／近年／近年では／ここ数年の間に］ ＋ ［見いだされている。／発見されている。／発見がなされている。／明らかにされている。／明らかにされてきている。／明らかにされつつある。／確認されている。／・／・］ ＋ 引用表示

が出現するならば、この文は「A12.既存の研究」であるが、

「筆者<r>は」 ＋ ［見いだしている。／発見している。／発見をしている。／明らかにしている。／確認している。／・／・］

が出現するならば、この文は「A24J.著者の立場（自己引用）」である。

第3B図 確定的ルールの例

準文献群とは別の新たな対象文献群として選択した，C型肝炎に関する日本語論文50件に適用した。これらのルールでは，③表層的パターン照合による表層的な言語解析によって，④文単位に順次ルールを適用した。

ルールでは，各カテゴリごとの出現確率や遷移確率を，当該文がそのカテゴリである可能性の高さを表わす得点とし，その文がルールの条件に合致するたびごとにそれぞれのカテゴリ別の得点を加算していく。たとえば，基準文献群における「…ている。」「最近」が出現する文が「A.12 既存の研究」である出現確率が，それぞれ 0.5, 0.3 である場合，対象文献群中のある文に「ている。」が出現したら，その文が「A12. 既存の研究」である得点は，0.5 であり，「最近」も出現していればさらに得点が 0.3 加算される。また，手がかり語の出現に基づくルールでは，上位カテゴリの確率を下位カテゴリに継承することも可能である。ルールの例を第３A図に示した。

なお，１文中に２つのカテゴリに相当する内容が含まれているものについては，以下の三つのパターンだけを扱った。

文の前半部のカテゴリ……文の後半部のカテゴリ

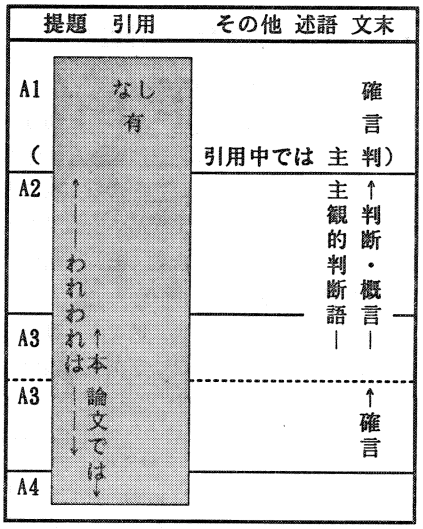
A11. 一般的見解……A21. 既存研究の問題点
A12. 既存の研究……A21. 既存研究の問題点
A312. 目的……A321. 研究方法の概要

ルールは数群に分け，対象文献群の１文目から文ごとに順次ルールを適用していき，ひとつのルール群が終了するごとにその文が持っている各カテゴリの得点を調べ，いずれかのカテゴリの得点が1.00を越えた時点でルールの適用を終了し，その時点で最も得点が高いカテゴリをその文に付与した。また，いずれのカテゴリの得点も1.00を超えなかった場合には，全ルールの適用が終了した時点でもっとも得点が高いカテゴリをその文に付与した。

この枠組みに基づいて，６群からなる146ルールを作成して50件のC型肝炎論文に付与したところ，全体の95.2%の文において，自動付与に成功した。

B. 複数領域の論文を対象とした場合の前報の枠組みの問題点

前報で用いたルールを用いて，今回新たに分析対象に追加した情報検索と対人認知領域の論文にカテゴリの自動付与を行なった。その結果，いずれの領域でも分析精度は５割未満であった。手がかり語の追加などによって，



第４図　分析の手がかり語の図

精度を少しは向上できると予測された。しかし，より大きく精度を向上させるためには枠組みの根本的な変更が必要であると考えた。

また，C型肝炎論文と，情報検索や対人認知領域の論文とは記述の仕方の特徴が異なる。すなわち，C型肝炎論文では，１文で一つの内容を述べ，一つのカテゴリが１文で完結する場合が多く，１文内の前半と後半でカテゴリが異なる文も多く見られた。それに対し，情報検索や対人認知領域の論文は，第２表に示したように，「A. 問題」カテゴリに相当する部分がC型肝炎論文より長く，連続した数文でひとまとまりの内容を述べ，１つのカテゴリが数文にわたって継続している記述の仕方が多く見られた。この「連続した数文でひとまとまりの内容を述べる」記述の仕方は，以下の理由により，自動付与において大きな問題である。

分析の手がかりは非常におおまかに示すと第４図のようになる。上記のような「連続した数文からなるひとまとまりの内容を述べている部分」中の２文め以降では，多くの場合，分析において重要な手がかりとなる「われわれは」や「本稿では」，あるいは，引用表示が省略される（図中網かけ部）。このような重要な手がかりとなる部分が省略された文だけを取りあげた場合，文末表現や述語からだけでは，「A1. 既知の事実」と「A3. 研究概要」の区別，著者自身の意見（「A2. 問題点の明確化」）と引用文献中で述べられている意見（「A12. 既存の研究」）の区別，「A11. 一般的見解」と「A12. 既存の研究」

の区別などがわからない。したがって，カテゴリを適切に付与するには，前の文とひとまとまりの内容を述べている続きであるかどうかを認定することが重要である。

したがって，C型肝炎論文だけを対象とする場合には，1文ずつ処理をし，1文より小さい部分に対してもカテゴリを付与するというように，1文より小さい単位に留意する必要があったのに対し，この3領域を同時に対象とする場合には，それに加えて，数文でひとまとまりの内容を述べている部分をひとまとまりのものとして認定するという，1文より大きい単位に留意する逆方向の処理も同時に行なう必要がある。

また，カテゴリの遷移確率の算出法も修正が必要である。というのは，第2表に示したように，情報検索や対人認知領域では，C型肝炎論文に比べ，カテゴリが繰り返し出現する複雑な出現型の論文が多かった。この場合，より大局的な出現型としてとらえれば類型化できるが，最も詳細なレベルでは次文へのカテゴリ遷移について一定の傾向は見られない。したがって，大局的な出現型にもとづいた遷移確率を用いたルールにする必要がある。

さらに，確率的なルールでは，カテゴリ遷移確率による得点に，個別の手がかり語の出現確率による得点を次々に積み重ねていく。そのため，個々のルールがどのように作用し，どのような効果を与えているかはわかりにくい。したがって，修正が必要になった場合，どのルールをどのように修正したり，削除したりすればよいかわからない。

これらの問題点を解決するため，自動付与の枠組みを以下のように変更した（第2図参照）。

C. 本報で用いた自動付与の枠組み

主要な変更点は，確定的なルールの導入と前文との連接関係の重視である。

前者については，別稿[2]の論議に基づき，文の主な述語の活用語幹と文末表現に着目し，それらと，前文のカテゴリ，接続表現，提題表現，引用の有無，副詞，その他の手がかり語などの条件の組合せを整理した。手がかり語は，別稿の論議に基づいて，意味や出現の仕方が類似しているものをグループ化した。そして，どの手がかり語（または手がかり語のグループ）がどのような条件の組合せで出現したら，必ずどのカテゴリになるということを決定できるパターンを選び，確定的なルールを作成した（第3B図参照）。

確定的ルールでカテゴリが決定しない場合には，確率的ルールを用いた。

また，連続した数文でひとまとまりの内容を述べている部分を認定するために，前文との連接関係を重視した。

連接関係は，別稿[2]で検討したように，接続表現，提題表現の繰り返し・省略・指示，列挙表現，述語の連鎖でとらえる。しかし，本稿では，表層的パターン照合という表層的な言語処理で，構文解析を行なっていないので格関係を扱うことができない。それを補うため具体的な助詞も含む形で言語表現のパターンを設定しており，そのために登録パターン数は多くなっているが，そのようにしても連鎖関係を的確にルールに反映するには限界がある。

一方，表層的なパターンを用いた分析は，構文解析や領域の専門的な知識も組み込んだ高度な言語処理に比べ，①ルールに手がかり語の組合せを直接記述するのでわかりやすく，修正が容易である，②領域を越えた適用が可能である，③処理時間が短い，④構文解析ルーチンなどの処理に必要な資源が少なくてすむ，という利点がある。②は情報メディアの伝達内容の特性を領域に関わらず共通にとらえるという構成要素カテゴリの基本的な目的に合致し，しかも，①は予備的な段階では特に重要であることから，本稿では表層的なパターン照合を採用した。

遷移確率については，上位レベルで大まかにとらえた出現型を用いた。しかし，これは，確定的ルールでカテゴリが決定しない場合に用いる確率的ルールの一部であるので，今後，確定的ルールの増強にともなって，重要性が低下する部分である。

IV. カテゴリ自動付与の結果と検討

A. 自動付与の実施

カテゴリ自動付与のルールを簡略化するために，ルールの適用に先だって，あらかじめ，分析対象には以下の前処理をした。いずれの処理も機械的に行ない，人手による修正は加えていない。自動付与のルールでは，これらの処理によって認定したグループを参照する。

① 接続表現の認定と接続表現グループの認定
② 述語の活用語幹の認定と述語グループの認定
③ 文頭にある前方指示表現を含む提題表現の認定

①では，接続表現を〈順接〉，〈逆接〉，〈添加〉，〈対比〉，

〈同列〉,〈転換〉,〈補足〉の7グループ分けたリストと分析対象中の各文を照合し，リスト中の接続表現を含む文には，その接続表現の出現位置と該当するグループの記号を付記した。

②も同様に，述語の活用語幹を意味や出現の仕方によって，〈存在〉,〈研究(実施)〉,〈成果〉,〈報告〉,〈思考〉などのグループに分けたリストを作成し，分析対象中の各文の文末にある述語の活用語幹のグループを調べた。

③に関しては，分析対象中の各文に，以下の前方指示表現を含む提題表現が，文頭，もしくは文頭の接続表現の次に出現するか調べた。

（1）「これは」,「前者については」など，指示語に提題助詞がついたもの

（2）前文中の語に「は」などの提題助詞がついたもの

（3）「この」などの指示語に，前文中の語と「は」などの提題助詞がついたもの

以上の前処理をした後，前節 III 章 C で述べた枠組みに基づいて作成したルールを，分析対象中の文に1文ずつ順次適用し，全ての文にカテゴリを付与した。

B. 自動付与の結果

カテゴリ自動付与の結果が，あらかじめ人手で付与してあるカテゴリと一致した場合，自動付与に成功したと判定した。文単位で見たカテゴリ自動付与の分析精度は，確定的なルールを追加するにつれ，第5図のように徐々に改善された。今後もさらにルールを増強することにより，分析精度は改善されると予測される。

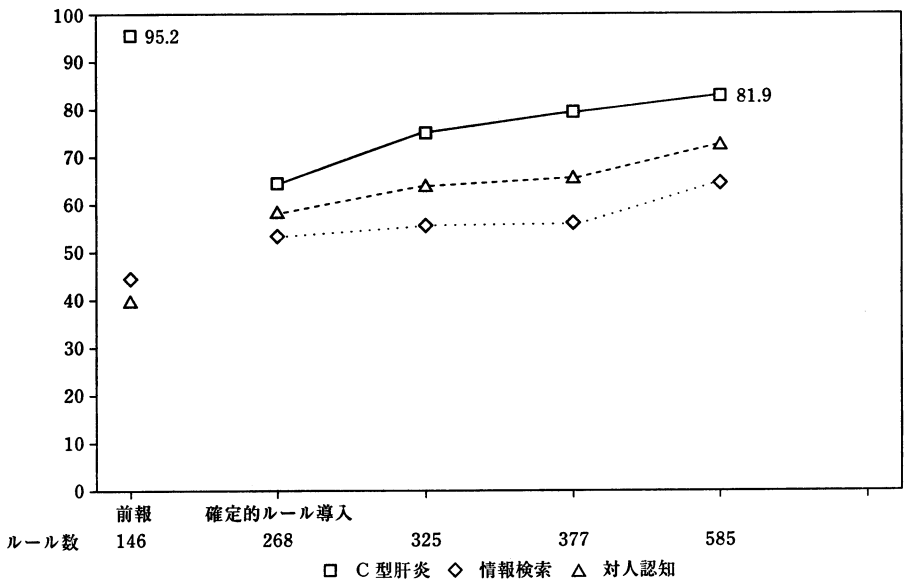
第3表　構成要素カテゴリ自動付与の精度（領域別）

	カテゴリが決定した文*	自動付与に成功した文	精度
C型肝炎	234 文	213 文	91.0%
情報検索	653	561	85.9%
対人認知	1037	903	87.1%

*: 確定的ルールによってカテゴリが決定した文と確率的ルールによる得点が 1.00 を超えてカテゴリが決定した文の合計

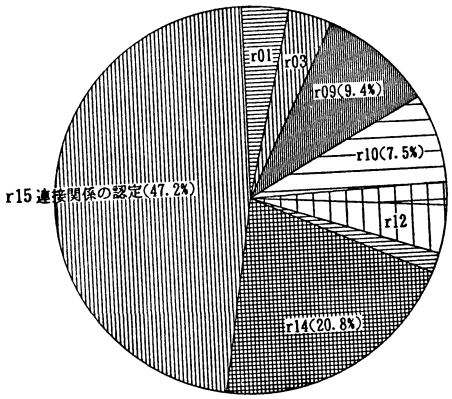
また，第5図の精度は，確定的ルールでカテゴリが決定せず，確率的なルールでもいずれのカテゴリの得点も1.00に満たなかった文についても，最終的に1.00未満でもその中で最も得点が高いカテゴリを付与した結果に基づいて算出している。それに対し，確定的なルールもしくは確率的なルールで得点が1.00を越えてカテゴリが決定したものだけを調べると，第3表のように，いずれの領域の対象論文においても，9割近い精度に達している。

本稿で自動付与に用いたルールは，第6図のように，15のルール群に分けられる。自動付与に失敗した原因を

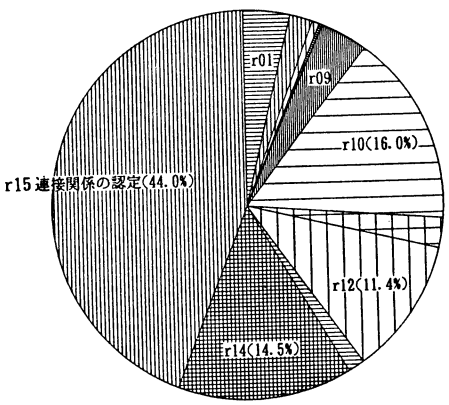


第5図　分析精度の推移

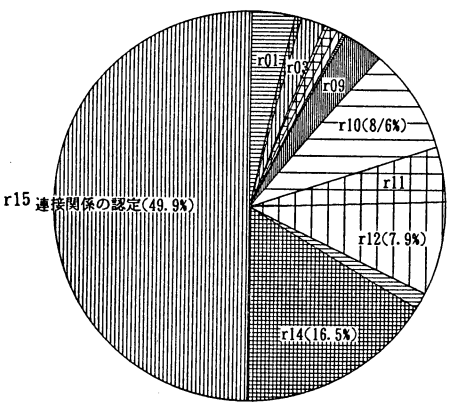
ルール群番号	当該ルール群が認定する主なカテゴリ
r01	前文と同じカテゴリ
r02	「A12. 既存の研究」
r03	「A24J. 自己引用」
r04	「A4. 用語の定義」
r05	「A311. 仮説」
r06	「A324. 論文の構成」
r07/r08	「A312. 目的」
r09	「A32. 研究範囲」
r10	「A22. 重要性・意義」
r11	「A21. 既存研究の不完全さ」
r12/r13	「A1. 既知の事実」
r14	「A24I. 著者の意見」
r15	前文との連接関係



C型肝炎



情報検索



対人認知

第6図 ルール別にみた自動付与の失敗

ルール群別に見ると，いずれの領域でも，自動付与に失敗したものの半数近くが連接関係の認定の失敗によるものであることがわかる。今後は，特に連接関係の認定に関わるルールを改善する必要がある。

C. 構造分析研究の比較

本稿と同様に，日本語で書かれた論文を対象として，言語的特徴に基づいて論文中の各文に何らかの役割表示を自動付与することによって論文の機能的な側面から見た構造を解析している研究がある[4)-6)]。

中本ら[5)]は，文献には，その内容をよく示す特徴的な「内容明示文」が含まれているとし，その「内容明示文」が文献中で果たしている役割として8種の「意味属性」を設定している。31論文中の「内容明示文」から抽出した特徴的な語句の構文パターンを用いて，パターン抽出に用いたのと同じ31論文に意味属性を付与し，さらに評価実験も行なっている。

三池ら[4)]は，全文データベースにおいて，対話型で，

第4表　機能構造分析研究の比較

		中本ら[5]（評価実験）		三池ら[4]	西村ら[6]	本稿
文の役割表示	呼称	意味属性		文役割	主題	構成要素ｶﾃｺﾞﾘ
	種類と例示	目的，方法，結果，背景，意見，特徴，課題，内容紹介 8種		話題，目的，背景特徴，結論，課題 6種	外部環境動向，従来研究の概要など（17種）	図1参照
	付与単位	文		文	文	文
	付与した文の割合	27.2%*	14.3%	約24%	−	全ての文
処理方式	特徴的な文	構文パターン		構文パターン	表層パターン	表層パターン
	1つの役割表示が継続する範囲の認定	2文以上からなる列挙構造は扱えない		文間の階層構造により文役割を複写	主題の範囲を明示する表層パターン、文の連接に関する言語的知識など	連接関係を考慮しているが、不十分
目的		ブラウジングのインタフェイスへの応用		全文検索での柔軟な処理戦略立案の支援	ブラウジングのインタフェイスへの応用	情報メディアの特性の記述（応用可能性も検討
分析対象	文献数	31文書	545文書	529文書	30文献	137文献
	掲載誌	東芝ﾚﾋﾞｭｰ	東芝レビュー	東芝レビュー	日本機械学会論文集／応用物理	3領域33誌
	1文献あたりの平均文数		69.9文	76.2文	−	（領域別に）73.6〜176.9文
	処理部分	文書全体	文書全体	文書全体	文書全体	「A．問題」部
	処理文数（構文解析成功数）	約1,500文*	38,098文（35,772文）	−	−	2,689文

*：中本ら[5]は、「全体文数に対して人手で抽出された内容明示文（「意味属性を持つ文（筆者注）」）の割合は、27.2%であった。」と述べているが、中本らが「表2－2　内容明示文数」で意味属性別に示している人手で抽出した内容明示文の数の合計は、1,015文であり、それを処理文数で除すと、処理文数全体から文の役割表示である「意味属性」が付与された文の割合となるはずであるが、それは約67.6%となってしまい数値が食い違ってしまっている。

利用者が入力した検索語に対して，「文役割」別にその語が当該役割の文に含まれる文献数の一覧を提示することによって，利用者の検索戦略の立案を支援することを意図している。文書の OCR 入力から，文書書式の解析，「文役割」の付与，文脈構造の解析，全文データベースの検索までの包括的なシステムを開発している。

西村ら[6]は，文の役割として，分野の課題，従来研究の概要，従来研究の問題点，本研究の問題点，今後の課題など「主題」を設定し，その「主題」を持つ文の特徴的な語句の表層的なパターンと「主題」のスコープを解析する知識を用いて「主題」を付与している。結果は例示的に示されており分析精度は不明である。

これらと本稿とを以下に比較し，本稿の特徴を明らかにし，改善の方向や課題を検討する（第4表参照）。なお，これらの研究はいずれも学会や研究会における報告であるため，詳細に記述されていない部分もある。

また，第4表に示すように，それぞれの研究では，文の役割を示すものの呼称が異なる。以下では，それらを一括して文の「役割表示」と称する。

1.　処理方式

a.　特徴的な文の役割の認定

付与する特徴的な文の役割表示の認定には，中本ら[5]と三池ら[4]は構文パターンを用い，西村ら[6]と本稿は，構文解析を行なわず，表層的なパターンを用いている。西村ら[6]は，形態素解析をした後，文頭と文末の数文節を用いてパターンを定義しており，表層的な解析手法を用いた理由として，①領域への依存性を少なくして移植性を高める，②解析部のブラックボックス化をさけて，ユーザ適応性を向上させる，③全文章の解析を現実的な時間内で行なうことをあげている。領域を越えた適用を志向している構成要素カテゴリでは①は重要である。しかし，表層的なパターンよりも構文パターンを用いた方がより少ないパターンで処理が可能であり，分析精度も改善されるという報告もある[11]。さらに，前文との連接関係を含め，分析基準を的確にルールに反映するには構文解析によって格関係を明確にする必要がある。構文パターンを用いている中本らや三池らがすでに500件以上の文書に対して自動付与を試みているという実績も考慮すると，今後は構文パターンの利用が有望だと考える。

b.　1つの文役割が継続する範囲の認定

連続した2文以上からなるひとまとまりの部分の認定

は，本稿において最も大きな問題となっていた。中本ら[5]は2文以上からなる列挙構造は扱えないが，三池ら[4]は，文脈構造解析によって2文以上からなる列挙構造等も扱えると述べている。西村ら[6]も「主題の範囲を明示する表層パターン」や「文の連接に関する言語的知識」などを用いていると述べているが，その具体的な内容は報告されていない。

したがって，三池ら[4]と西村ら[6]では，(1) 言語表現上の手がかり語を組み合わせたパターンによって，特定の役割表示を持った特徴的な文の認定と，(2) その役割表示が継続する範囲の認定という2種類の処理を用いている。本稿は，(2) の処理に関して，連接関係を重視することによってひとまとまりの部分の認定や役割表示の継続範囲の認定をしようとしたが，第6図からも明らかなように，自動付与に失敗した原因の半数近くが連接関係の認定処理の失敗によるものであった。すなわち，本稿では，(1) の特徴的な文の役割表示の認定にはある程度成功しているが，(2) の一つの役割表示が継続する範囲の認定に失敗している。

しかし，それぞれの研究は，以下に述べるように，付与する文の割合，分析対象文書などの条件が異なるので，単純に分析結果を比較することはできない。

2. 付与する文の割合

中本ら[5]と三池ら[4]は，各論文中の1〜3割程度の文だけに役割表示を付与している。それに対し，本稿では，全ての文に役割表示を付与し，しかも必要に応じて一文より小さい部分にも役割表示を付与するという違いがある。

これは，研究目的が異なるためであると考えられる。すなわち，中本ら[5]と西村ら[6]は，一つの文献の内容を文の役割に従って拾い読みできるインタフェイスの開発への応用を意図している。このような拾い読みや検索結果の絞り込み[4]という目的のもとでは，むしろ，一部の文だけに役割表示が付与されている方が有用であろう。

それに対し，本稿は，情報メディアの伝達内容の特性をとらえる一連の研究の一環として構造を分析しているため，全ての文にカテゴリを付与している。特定の応用のためだけに分析しているわけではないが，文献内の関係と文献間で共通の関係を保持しながら，文献内の部分に着目するという構成要素カテゴリの特性を利用した様々な応用が可能であると想定し，その一部を試みている[7]。

3. 分析対象

中本ら[5]と三池ら[4]は，500件以上の文献を処理しているが，いずれも「東芝レビュー」誌の掲載文献だけを対象としている。西村ら[6]は2誌の掲載論文を対象とし，本稿では，3領域の33誌に掲載された論文を対象としている。本稿や既報[1]の分析から，領域や掲載雑誌によっても記述の仕方の特徴が異なるので，それぞれの研究において分析対象の数を増やすとともに，分析対象の掲載誌や領域の拡大が望まれる。

処理の対象部分については，本稿では，今回は「A. 問題」カテゴリの部分のみに関して報告しているが，他の三つの研究は文書全体を処理対象としている。

一方，対象文書の長さをみると，中本ら[5]，三池ら[4]，本稿におけるC型肝炎論文は一論文あたりの平均の文数がほぼ同じである。本稿では，C型肝炎論文では，ひとまとまりの内容を述べている部分や一つの役割表示が継続している範囲を認定するという問題がほとんどなく，分析精度も高かった。それに対し，情報検索や対人認知研究の論文は長く，ひとまとまりの内容を述べている部分が多く，そのようなひとまとまりの部分の中の文には分析の手がかりとなる特徴的な語があまり出現しないため，このような部分に対する役割表示の付与が大きな問題となっていた。

このような一つの文の役割表示が継続する範囲の認定には，三池ら[4]の方式が有用であると思われるが，三池らと本稿とでは，対象文書の長さや文役割を全文に付与するかどうかという処理条件の差異を考慮しなければならないだろう。

4. 分析効率

中本ら[5]は，特徴的な語句の構文パターンを用いて，その構文パターンを抽出したのと同一の文献群に対して「意味属性」を自動付与したところ，自動付与した文のうち，正しく付与できた文の割合（精度）を示す「抽出成功率」は，意味属性別にみて 69.9%〜97.7% であった。次に，評価実験として，「東芝レビュー」誌の545論文に自動的に意味属性を付与したところ，「抽出成功率」は，意味属性別にみて 43%〜98%，平均 81% であった。また，人手で「意味属性」を付与した文のうち，自動付与でも認定された文の割合（再現率）を示す「カバー率」は 75% であった。

本稿は，「A. 問題」カテゴリの部分だけについて報告し，しかも全ての文に文役割を示すカテゴリを付与して

いるので，一概に比較することはできないが，分析精度は，領域別に見ると，全体で 65～82%，カテゴリが決定した文においては 86～91% であった。

以上，論文中の各文に文の役割表示を付与している諸研究を比較検討したが，その目的や付与の範囲などは様々であった。しかし，処理方式は，(1) 言語表現の組合せからなるパターンによる特定の文の役割表示を持つ特徴的な文の認定と，(2) 一つの文役割表示が継続する範囲の認定という 2 種類の異なる処理からなると考えられる。

本稿では，一部の役割を持つ部分だけに対する処理について報告したが，失敗の原因は主として (2) の処理にあった。しかも全ての文にカテゴリを付与し，長い論文も対象に加えたことによって，この失敗が強調された。今後は言語表現のパターンを追加することによって分析精度は向上すると予想される。しかし，既存の研究から，表層パターンよりも構文パターンの方が少数のパターンで効率よい処理が可能であることが示唆され，分析基準を的確にルールに反映するためにも構文パターンの方が望ましいと考える。その際，いままで蓄積してきた表層的な言語表現パターンを構文パターンに変換して活用することができると考える。

V. 自動付与の今後の課題

構成要素カテゴリの自動付与について，今後の展開を展望するときに，以下の点が問題である。

① 連接関係と一つのカテゴリの継続範囲の認定
② 全体的処理戦略の導入
③ 内容判断を伴うカテゴリの処理

まず，①の連接関係の認定は，前節における他の研究との比較においても本稿の処理における失敗の主要な原因として指摘した部分である。なお，実際の論文における記述の仕方は様々であり，一文ごとにカテゴリがかわっていく部分もあるので，ただ，文が連続している，あるいは同一段落中だからという理由だけで前文と同一カテゴリとは認定することはできず，文間の関係に基づいて一つのカテゴリが継続する範囲を認定する必要がある。

そのためには，直前の文だけでなく，より前方の文や段落，あるいは後方の文との関係も考慮し，連鎖関係をとらえるための言語表現のパターンを追加する必要がある。提題表現の前方照応に関しては，先行する語と同義語・類義語や上位下位関係にある語による言い替えや語句の一部分の省略にも対応するには，語義の管理も必要になる。このような語の意味に関する知識を用いると領域をこえた一般化が困難になるという問題点があり，一般化を優先するか，処理の精度を優先するかという方針の検討が必要である。

つぎに，②の全体的処理戦略に関しては，従来から，構成要素カテゴリを用いた分析では，論文全体からみた役割と局所的な役割とを二重に付与する「入れ子構造」を認めている。全体的な視野がないとこのような構造は認定できない。

一方，van de Velde[12] によると，人間がテキストを理解するには，分析的 (analytical)，連続的 (sequential)，全体的 (wholistic) という 3 つの戦略があり，これらは，それぞれボトムアップ処理，オンライン処理，トップダウン処理に相当する。構成要素カテゴリは，そもそも，論文中のある部分が論文全体から見てどのような役割を果たしているかを示し，全体的な視野の中で考えられるべきものであり，全体的処理戦略，すなわちトップダウンな処理に関わるものである。

それにも関わらず，現在の分析枠組みでは，手がかり語の組み合わせという分析的，すなわちボトムアップな方法によって，論文の初めから順次一文づつ処理をするという連続的な方法を採っているという矛盾がある。したがって，構成要素カテゴリの本来の特性を考慮すると，分析枠組みには「全体的な処理戦略」を取り入れる必要がある。具体的には，特徴の明確な部分を認定し，おおまかに全体の構造をとらえた上で，それらの特徴的な文との前後のつながりを見ながら細かい関係を認定していくことになるだろう。このような処理戦略の導入は，①のカテゴリの継続範囲の認定にも有用であると考える。

さらに，③内容判断が関わるカテゴリの扱いに関しては，現在の構成要素カテゴリの中には，「A22. 重要性・意義」や「A323. 結果・結論」のように，手がかり語の組み合わせからなる表層的な言語パターンだけでは区別が付かず，文の内容を評価してカテゴリを付与しなければならないものが含まれている。例えば，「A323. 結果・結論」は，その論文で取りあげる研究課題を述べる際に，研究の概要を示す一環として主要な結果や結論を記述している部分を表わすカテゴリである。

これは，第 7 図 b) のように具体的な結果が記述され

a) そこで著者らはこの手法を試みた結果興味ある成績を得たので報告する。

具体的な内容なし→A321. 研究方法の概要

A321. 研究方法の概要

b) そこで著者らは血液疾患におけるHCV抗体陽性率と輸血歴、肝疾患との関連について検討した。輸血歴のある血液疾患患者でHCV抗体陽性率が高率であり、慢性肝疾患が多いことが確認され、一方輸血歴のない血液疾患患者でも一般献血車に比べHCV抗体陽性率が高値であることが判明したので報告する。 A323. 結果

c) 著者らは、この流行性肝炎の実態を再調査するとともに、罹患した患者の追跡調査を行った。その結果、この流行性肝炎がHCVによることを示唆する成績を得たので報告する。 →A323. 結果？A321. 研究方法の概要 ？

第7図 内容判断を伴うカテゴリの例

ている文には「A323. 結果・結論」のカテゴリ付与する。a) のように，得られた結果の具体的内容が記述されていないものは，「A323. 結果・結論」とはせず，「A321. 研究方法の概要」とする。しかし，c) は「この流行性肝炎が HCV によることを示唆する成績」が具体的な結果といえるかどうかは人によって判断が分かれるであろう。しかも，これらは，いずれも，「著者らは…した結果…成績を得たので報告する。」という表層パターンをもっており，言語表現の表層的なパターンからだけでは両者を区別することは難しい。分析者による内容の判断に依存する場合は，基準が明確でなく，分析者によって分析結果が異なる可能性がある。

このような内容の判断によってカテゴリ付与が決まるものは，いずれも，一つ上位のレベルで考えれば同一のカテゴリである。したがって，これらを認定する明確な基準をさらに検討し，もし明確な基準が得られない場合は，分析の一貫性を保つ上から，これらのカテゴリを削除してより上位のカテゴリでまとめることも考える必要がある。

VI. ま と め

本稿では，複数領域の原著論文に対する構成要素カテゴリの自動付与について検討した。

すでに，人手で全文に構成要素カテゴリが付与してあるC型肝炎，情報検索，対人認知領域の論文を分析対象とした。これらの分析対象論文では，論文の長さ，カテゴリの出現順や繰り返しの程度，対象となる章の数や章の見出し，記述の仕方は様々であった。これは，領域ごとの記述の特徴の差異によるカテゴリ自動付与の問題点を調べる上で，適当なサンプルであると思われる。

他方，このように，論文ごとに長さや表面的な論文の構成が異なっていても，カテゴリによって特定の機能を果たしている部分を共通に指定できることは，構成要素カテゴリの利点の一つである。

上述の3領域の論文の「A. 問題」カテゴリの部分に対して，構成要素カテゴリの自動付与を試みた経過を報告した。その結果，自動付与失敗の主要な原因は一つのカテゴリが継続する範囲の認定の失敗であり，今後の展開を考えたときに，処理の枠組みに全体的な処理戦略を取り入れる必要性が示された。また，今後は，表層的なパターンと構文パターンのどちらを用いるか，語の意味に関する知識を用いるか，内容判断が必要なカテゴリの扱いに関して検討する必要があることが示唆された。それに合わせて，適用範囲の拡大，応用の実現とその有用性の検討も必要であると考える。

本稿をまとめるにあたり，ご指導いただきました慶應義塾大学文学部の上田修一教授に心からなる謝意を表します。

本研究は，平成5年度文部省科学研究費補助金（特別研究員奨励費）による成果の一部である。

1) 神門典子．構成要素カテゴリーを用いた情報メディア内部構造分析の試み：原著論文を例として．東京，慶應義塾大学，227 p，1991．平成2年度修士論文．(分析結果の概要は，神門典子．構成要素カテゴリを用いた原著論文の内部構造分析．情報処理学会研究報告（情報学基礎研究会 92-FI-25)). Vol. 92, No. 32, p. 39-46 (1992) 参照)
2) 神門典子．構成要素カテゴリを用いた情報メディアの構造の分析：言語表現に関する考察に基づく分析基準の再検討．Library and Information Science. No. 31, p. 39-49 掲載予定
3) 神門典子．情報メディアの構造：伝達内容の分析と

利用. Library and Information Science. No. 30, p. 1–19 (1992)
4) 三池誠司，小野顕司，住田一男．文書の構造解析に基づく文書情報検索．情報処理学会研究報告（情報学基礎研究会 93–FI–31). Vol. 93, No. 78, p. 39–46 (1993)
5) 中本幸夫，野上謙一，矢島真人，田野崎康雄．文書への意味属性付与のための意味辞書の拡張．情報処理学会第 45 回全国大会講演論文集（第 3 分冊）. p. 211–212 (1992)
6) 西村健士，島津秀雄．特定表現の重点的解析による科学技術論文構造化手法．情報処理学会研究報告（情報学基礎研究会 93–FI–29). Vol. 93, No. 39, p. 35–42 (1993)
7) 神門典子．原著論文の機能構造の分析とその応用：C型肝炎論文を対象とした基本動向記述文の抽出とその前提としての構成要素カテゴリ自動付与の試み．図書館学会年報．Vol. 40, No. 1, p. 11–20 (1994) 掲載予定
8) Swales, John; Najjar, Hazem. The writing of research article introductions. Written Communication. Vol. 4, No. 2, p. 175–191 (1987)
9) Crookes, G. Towards a validated analysis of scientific text structure. Applied Linguistics. Vol. 7, No. 1, p. 57–70 (1986)
10) 倉田敬子，神門典子．索引作成過程において索引作成者が用いる認知的枠組み．1993 年度三田図書館・情報学会研究大会，東京，1993 年 10 月．
11) 矢島真人，栗原基，岩井勇，鈴木謙二，田野崎康雄，近藤隆志．文書への意味属性付与のための意味辞書の開発．情報処理学会第 43 回全国大会講演論文集（分冊 3）. p. 325–326 (1991)
12) van de Velde, Roger G. Text and Thinking: on Some Roles of Thinking in Text Interpretation. Berlin, Walter de Gruyter, 1992, 328 p. (Research in Text Theory; Vol. 18)