情　報　学　基　礎 33－4
（1994．5．18）

文章特性が異なる複数領域の原著論文の構造

神門　典子

日本学術振興会特別研究員

C型肝炎、情報検索、対人認知領域の日本語原著論文計137件を対象として、構成要素カテゴリの自動付与を行なった。構成要素カテゴリは文献の内容の特性を主題領域に関わらず共通にとらえる枠組みとして提案したものである。分析に先立ち、学術的な文献の利用特性とそれから導かれる分析の要件について論じた。手がかり語の組合せパターンとそれを補う確率的なルールによってカテゴリを自動付与し、カテゴリが決定した文における分析精度は領域別に86～91%であった。自動付与の失敗原因の分析、対象文献の文章特性の検討、関連研究の検討から、(1)文間の連接関係の重視、(2)構文パターンの採用、(3)全体的処理戦略の導入、(4)内容判断を必要とするカテゴリの検討の必要性が示唆された。

Structure of Research Articles with Various Text Features

Noriko KANDO

Research Fellow of the Japan Society for the Promotion of Science

Graduate School of Library and Information Science, Keio University

2- 15- 45, Mita, Minato- ku, Tokyo 108, JAPAN

E- mail : *noriko@slis.flet.mita.keio.ac.jp*

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. Firstly, characteristics of academic literature usage and requirement of analysis are discussed. Automatic assignment of the Categories is conducted to the corpus of research articles with various text features by lexical- clue- combination- patterns matching and probabilistic rules. The result is, 86%- 91% of Category fixed sentences are assinged correct Categories automatically. And through analysis of errors, related researches and text features, followings are suggested to improve precision 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.

1．原著論文の構造解析

筆者は、日本語で書かれた原著論文を対象として、伝達内容の特性をとらえるために、伝達内容の構造を分析してきた[1-4]。

ここでは、一連の研究の基本的な考えとして、利用特性から導かれる要件を再確認し、関連研究との比較を通じて立場を明確する。そのうえで複数領域を対象とした分析について報告する。

1.1.構造解析に対する基本的なアイディア

伝達内容の構造をとらえる基本的な要件として、①文献の部分を単位として分析的に、②文献内の部分間の関係を保持し、③文献の種類に応じて文献間で共通な枠組みに基づいて、文献の伝達内容をとらえることとする。

一方、学術的な文献は、ただ読むだけでなく、新たな文献の生産や問題解決のために利用される。上記の要件は以下の学術的な文献の特性と密接に関わっている。

ｱ) 送り手と受け手が基本的に同領域に属している。
ｲ) 部分だけ読む場合がある。
ｳ) 内容の質や妥当性を検討するために、文献内の他の部分との整合性に着目して読む。
ｴ) 複数文献の内容を比較、統合して利用する。それが、情報の再生産、問題解決に通じる。
ｵ) 多数の文献を対象とした蓄積と検索を行なう。

利用特性からみると、ｲ)から①が、ｳ)から②が、ｴ)とｵ)から③が必要である。一方、ｱ)から、種々の慣習や規範が共有され、特に原著論文では、投稿規定や編集過程によって、役割や機能からみた内容構成に関する慣習や規範が強められていると想定される。

1.2.構造の種類

ここで扱うのはレイアウトなどの物理的な構造やＳＧＭＬで扱う章、節、段落という文書書式の論理的な構造ではない。伝達内容の構造には、修辞(意味)構造と機能構造とがある。前者は話題の連鎖、文や要素間の連接関係（順接、逆接など）などから捉え、後者は文献中の各部分が文献中で果たしている役割や機能という面から捉える。

②や文献の要約・理解には修辞構造と機能構造の双方が関わるが、③の文献間の共通性という要件も考慮し、機能面から捉えた構造に着目した。

1.3.構成要素カテゴリの設定

原著論文の構造を分析することの実現可能性を検討し、その意義、応用の可能性や有用性を検討するための具体例として、構成要素カテゴリを提案した[1)]。各種の論文執筆マニュアルと実際の原著論文の分析を通じて、一連の構成要素カテゴリを設定した。各カテゴリは論文全体からみた当該部分の役割や機能を表わす。

論文中の全ての文にカテゴリを付与し、カテゴリ出現順によって機能面からみた構造を表わす。必要に応じて、１文の前半と後半に異なるカテゴリの付与や、全体からみた役割と局所的役割とを二重に付与する「入れ子構造」も認めている。

複数分野の日本語原著論文を対象として人手で分析し、共通のカテゴリを付与でき、論文の長さや章節の見出しが様々でも共通のカテゴリで内容を捉えられた。カテゴリ毎に特徴的な手がかり語が認められ、カテゴリ出現型は類型化できた[1-2)]。

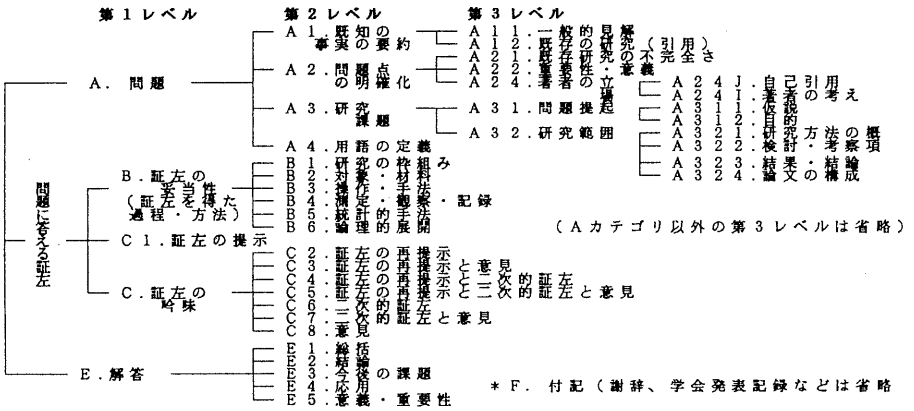

図１　原著論文を分析するために設定した構成要素カテゴリ

妥当性を検証するため、複数分析者間の一貫性を調べ、分析基準の問題点の解決を図った[3]。カテゴリ自動付与は、応用の実用化の前提として必要であり、分析基準明確化の面でも重要である。C型肝炎論文の「A. 問題」の下位カテゴリに対して、手がかり語の出現確率とカテゴリの遷移確率に基づくルールを用いてカテゴリ自動付与を予備的に試み、95.2%の精度を得た[4]。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[1),5-6]。そこで、本稿では、文章特性が異なる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、その問題点を検討する。

1.4.関連研究との関係

日本語論文に文の役割表示を付与し、機能面からみた構造を自動解析している研究としては、全文検索システムの開発[7-8]や1文献の内容を拾い読みするためのインターフェイス開発[9-10]を意図したものがある。いずれも1～2誌の雑誌の掲載記事を対象とし、本文中の1～3割程度の特徴的な文に役割表示を付与している。

これに対し、本研究での自動付与は予備的な検討であるが、①特定の応用だけを目的としない、②論文中の全ての文にカテゴリを付与し、③複数の雑誌や領域を対象とする方針をとっている。

①目的と応用： 本稿は論文の伝達内容の特性を捉えることを目的とし、特定の応用だけを志向していない。しかし、1.1.の要件により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て全文データベース検索や主題分析など様々な応用が可能であると想定している[1)]。特に、大量の文献群から必要な部分を直接検索し、必要に応じて機能面からみた部分間の関係に従って、文献内の関連部分を参照したり、他文献中の関連部分を横断的に通覧するなど、従来の文献単位の方式より柔軟な部分単位の検索・提示方式への応用可能性を想定し、その一部を試みている[4)]。

また、機能構造の知識があると論文内容をよく理解、記憶できるという報告もある[11)]。索引作成者のプロトコルでも、特定役割の文への着目が示された[12)]。このことから、機能構造は規範や慣習として文献の生産に関わっているだけでなく、文献の読み方にも関わっていると考える。

②付与する文の割合： 1文献の内容を拾い読みして直感的に概要を把握する[9-10]、あるいは文献単位での検索結果の絞り込む[7-8]ために文の役割表示を用いるには、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な場合もある。

それに対し、本稿は特定の応用は意図せず、学術文献の利用特性を考慮し、1.1.ｳ)の文脈に沿った妥当性の検討も重要であるという立場から全文を対象としている。文献中では必ずしも関連部分が連続して出現するわけではなく、特徴的な文の前後を見るだけでは文脈を把握できない。たとえば、実験方法についての検討は、実験方法に関する記述の前や後ろ、結果の記述の前や後ろ、考察部分など様々な位置に出現し[1)]、方法を記述した文の前後を見るだけではわからない場合が多い。

③対象文献： 人手での分析結果[1)]から、領域によって、あるいは論文の長さ・研究課題の種類・掲載雑誌などによって、論文の長さ・文章の特性・出現語句・章節の見出しなどが異なる。投稿規定や編集の過程や、読者と著者の重複から、雑誌毎にある傾向が生じる可能性がある。

2. 複数分野を対象とした構造の自動的な分析

2.1.分析対象: 論文の中心課題を示すまでの「A. 問題」カテゴリを対象とした。ここはいわゆる主題文を含み、主題分析や自動索引への応用、領域の現状把握[4)]等の種々の応用が考えられる。

表1 分析対象論文

	論文数	論文全体		A.問題ｶﾃｺﾞﾘ	
		1件ｱﾀﾘ平均		総数	1件ｱﾀﾘ平均
	(件)	(文)*		(文)*	(文)
C型肝炎	50	73.6		293	5.9
情報検索	49	193.4		914	18.7
対人認知	38	176.9		1,482	39.0

*:本文(章節の見出しは除く)。句点「。」で区切られたものを1文とする。箇条書きでは、句点がなくても文とした。

2.2.確率型ルールを用いたカテゴリ自動付与

前報[4)]の手がかり語の出現確率とカテゴリの遷

移確率に基づく確率型のルールを用いてカテゴリを自動付与した。１文ずつ順次処理し、１ルール群が終了する毎に各カテゴリの得点を評価し、いずれかのカテゴリの得点が1.00を超えた時点でルールを終了し、最も得点が高いカテゴリをその文に付与した。全ルール(146)の適用終了後、いずれのカテゴリの得点も1.00を超えない場合は、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。精度は、Ｃ型肝炎では95.2%であったが、検索と対人認知ではいずれも５割未満であった。

2.3.分析対象論文に見られた文章特性の差異

情報検索や対人認知論文とＣ型肝炎論文とでは、用語や論文の長さ、１文の長さ、多く出現するカテゴリの種類、引用した文献内容の記述法などが異なるだけでなく、表２のごとく文章の記述の仕方自体にも差が見られ、これが自動付与失敗の原因と考えられた。特に、情報検索や対人認知論文で多く見られた「連続した数文でひとまとまりの内容を述べている部分」が自動付与では問題となる。このような部分中の２文め以降では、重要な手がかりとなる語句や引用表示が省略され、その文だけを取りあげた場合、残りの語句からだけでは適切にカテゴリを判断できない。適切にカテゴリを付与するには、前文とひとまとまりの内容を述べている続きであることを認定する必要がある。

表２　文章の記述のしかたの特性

	Ｃ型肝炎	情報検索・対人認知
「A.問題」ｶﾃｺﾞﾘ部	「A.問題」が短い １文ずつ完結	「A.問題」が長い 数文ひとまとまり
1ｶﾃｺﾞﾘの継続範囲	1文または1文未満 1文内の前後半でｶﾃｺﾞﾘが異なる場合もある	1～数文
分析の留意点	文以下の部分	文以上のまとまり

Ｃ型肝炎論文だけの処理では、１文ずつ処理し、１文より小さい単位に留意する必要があった。３領域を同時に処理するには、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という１文より大きい単位にも留意するという逆方向の処理を同時に行なう必要が生じる。

また、カテゴリ遷移確率の算出法も修正が必要である。表３のように、情報検索や対人認知の論文は、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が多かったからである。

表３　論文ごとに見た構成要素カテゴリの出現型（第２レベル）*

	肝炎	検索	認知(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

*：論文内でカテゴリが出現した順に左から記述。（　）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。

さらに、確率的ルールでは、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ため、個々のルールの作用や効果、失敗の原因、問題点、修正が必要な箇所がわかりにくい。これらの問題点を解決するため、以下のように枠組みを変更した。

2.4.新しい自動付与の枠組み

①確定的ルールの導入：　言語表現の面の検討に基づいて[3)]、主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。21群633語の手がかり語を認定し、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。カテゴリを確定できる手がかり語（グループ）の組合せパターンを抽出し、それに基づいて確定的なルールを作成した。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、今回は予備的な検討であるため、処理時間の速さなどを考慮し、表層的な言語処理とした。

②連接関係の重視：　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、接続表現、提題表現の繰返し・省略・指示、列挙表現、述語の連鎖でとらえる[3)]。しかし、本稿では、表層的な言語処理のため、連鎖関係を的確にルールに反映するには限界がある。

③遷移確率：　上位レベルで大局的にみた出現

型を用いた。

表層的な言語処理によって、1文ずつ順次処理を行なう点は前報と同じである。

3．カテゴリ自動付与の結果

自動付与の結果が、あらかじめ人手で付与したカテゴリと一致した場合、成功と判定した。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て徐々に改善されつつあり、585個のルール[脚注]で、C型肝炎で82%、情報検索で65%、対人認知で73%であった。

自動付与の失敗の原因をルール群別に見ると、いずれの領域でも失敗の半数近くが連接関係が正しく認定できないためであった（表4）。今後は連接関係の認定のルールの改善が必要である。

表4　ルール群別に見た自動付与失敗

ルール群番号	当該ルール群が認定する主なｶﾃｺﾞﾘ	(%) 肝炎	(%) 検索	(%) 認知
02	「A12.既存の研究」	0	0	0.5
03	「A24J.自己引用」	3.8	1.8	2.0
04	「A4.用語の定義」	0	0	0.2
05	「A311.仮説」	0	0	0.7
06	「A324.論文の構成」	0	0.6	0.5
07/08	「A312.目的」	0	0.3	0.2
09	「A32.研究範囲」	9.4	3.7	3.5
10	「A22.重要性・意義」	7.5	16.0	8.6
11	「A21.既存研究の不完全さ」	1.9	2.5	4.4
12/13	「A1.既知の事実」	5.7	12.6	9.1
14	「A241.著者の意見」	20.8	14.5	16.5
01/15	前文との連接関係	51.0	48.0	53.6

カテゴリが決定した文、すなわち特徴が比較的捉えやすい文だけをみると、いずれの領域でも精度は平均86%以上であった（表5）。

表5　カテゴリが決定した文における精度

	肝炎	検索	認知
カテゴリが決定した文＊	234	653	1037
自動付与に成功した文	213	561	903
特徴的な文での精度(平均)	91%	86%	87%

＊: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

4．関連研究と比較

1.4.で述べた諸研究を概観し改善の方向や課題を検討する（表6）。

4.1.処理方式

①特徴的な文の役割の認定：　中本らと三池らは構文パターンを用い、西村らは構文解析を行なわず、形態素解析後、文頭と文末の数文節を用いて定義した表層的パターンを用いている。

西村ら[10]は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。

しかし、構文パターンの方が表層パターンより少ないパターン数で処理可能であり、分析精度も改善されるという報告がある[13]。しかも、前文との連接関係の認定や、分析基準の的確なルールへの反映には格関係の明確化が必要であることから、構文パターンが有望だと考える。

ただし、本稿で対象とした中では、1文が800字を超える長い文や体言止めの文も少なくなかった。これらの文で構文解析がどの程度成功するかが処理の有効性を決めるポイントとなるだろう。

②1つの文役割の継続範囲の認定：　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、本稿で最大の問題であった。

三池らと西村らは、①手がかりとなる言語表現のパターンによる、特定の役割を持つ特徴的な文の認定と、②その文役割が継続する範囲の認定という2種の処理を行なっている。

本稿は、①の特徴的な文の役割の認定はある程度成功したが、②の継続範囲の認定は、文の連接関係の認定で対応しようとしたが不充分であり、失敗の原因は主として②であった。

文役割の継続範囲の認定には、修辞構造も関わり、構文解析だけでなく、語の類義、上位下位関係などの意味関係が必要である。しかし、意味関係を扱うと領域依存性が高まるという問題がある。

4.2.自動付与処理の条件

各研究は、1.4.で述べたように、研究目的、全文に付与するかどうか、分析対象文書の件数・長

脚注:ルールを簡略化するため、以下の前処理をした。いずれも自動的に処理した。ルールはここで認定したグループを参照することができる。① 接続表現と接続表現グループの認定、② 述語の活用語幹と述語グループの認定、③ 提題表現の認定（文頭の前方指示表現を含む）

さ・掲載誌の種類・領域などの条件が異なり単純に比較できない（表６）。

本稿ではカテゴリ出現型が単純な短い論文（Ｃ型肝炎）やカテゴリが決定した特徴を捉えやすい文では精度が比較的よかった。しかし、(1)論文中の全ての文にカテゴリを付与し、(2)長い論文も対象とするため、ひとまとまりの内容を述べている部分に多く見られる特徴的手がかり語が少ない文の処理も必要であった。

５．今後の課題

本稿は予備的な分析であるが、カテゴリ自動付与の実行可能性が示された。また、手がかり語パターンの追加によりさらに精度の向上が見込めるが、以下のような改善の方向も示された。

①連接関係とカテゴリ継続範囲の認定の強化

②全体的(Wholistic)な処理戦略の導入

③内容判断を伴うカテゴリの処理の検討

また、原著論文や学術的な文献の特性を踏まえ、他の構造の分析との関わりや応用の実現と有用性の検討、規範や慣習の経時的検討が必要であろう。

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上田修一教授に謝意を表します。

表６　機能構造分析研究の比較

	中本ら9) 実験１*	実験２*	三池ら7)**	西村ら10)	本稿
目　的	ブラウジングのインタフェイス		全文DBで柔軟な検索戦略立案の支援	ブラウジングのインタフェイス	文献の特性の把握（応用可能性も検討）
文の役表示	意味属性		文役割	主題	構成要素ｶﾃｺﾞﾘ
役割表示の種類	目的,方法,結果,背景,意見,特徴,課題,内容紹介		話題,目的,背景特徴,結論,課題	外部環境動向,従来研究の概要など	図１参照
付与する単位	文		文	文	文
付与した文の割合	27.2%*	14.3%	約24%	－	全ての文
分析対象文献数	３１	５４５	５２９	３０	１３７
対象文献掲載誌	東芝レビュー	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文献の長さ（平均）		69.9文	76.2文	－	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
特徴的な文の認定方式	構文パターン		構文パターン	表層パターン	表層パターン
役割の継続範囲認定	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ﾊﾟﾀｰﾝ 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
分析精度	抽出成功率（意味属性別に）69.9～97.7% 43～98%（平均81%） カﾊﾞｰ率 75%				（領域別に）全文平均 65～82% ｶﾃｺﾞﾘ決定文平均86～91%

*実験１はパターン抽出に用いた文献群に対して、実験２は評価実験として別の文献群に対して行なった。
**三池らはOCRによる文書の読み込みから全文検索までの総合的なシステムを開発している。　***内訳は表１参照

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．慶大，227p，1991．修士論文．
2) 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(92-FI-25)．Vol.92, No.32, p.39-46(1992)
3) 神門典子．構成要素カテゴリの分析基準の再検討：言語表現の考察に基づいて．Library and Information Science. No.31, p.39-49掲載予定
4) 神門典子．構成要素カテゴリを用いた特定領域の研究動向把握の試み：Ｃ型肝炎論文を対象とした構成要素カテゴリ自動付与研究の応用として．1992年度三田図書館・情報学会研究大会，1992年．
5) Swales, J.et al. The writing of research article introductions. Written Communication, Vol.4, No.2, p.175-191(1987)
6) Crookes, G. Towards a validated analysis of scientific text structure. Applied Linguistics. Vol.7, No.1, p.57-70(1986)
7) 三池誠司ほか．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(93-FI-31)．Vol. 93, No.78, p.39-46(1993)
8) 三池誠司ほか．自動抄録機能をもつ対話的文書検索システム：検索機能．情報処理学会第48回全国大会講演論文集(第３分冊)．p.277-278(1992)
9) 中本幸夫ほか．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)
10) 西村健士ほか．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(93-FI-29)．Vol. 93, No.39, p.35-42(1993)
11) Samuels S. et al. Adults' use of text structure in the recall of a scientific journal article. J.of Educ. Res. Vol.81, No.3, p.171-174(1988)
12) 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．
13) 矢島真人ほか．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（第３分冊）．p.3:325-326(1991)