自然言語処理 70－2
（1989. 1. 20）

文脈構造の分析*

小野 顕司　浮田 輝彦　天野 真家

（株）東芝 総合研究所

本稿では日本語テキスト，特に論説文を対象にして文章の全体の構造を分析する．文章全体に含まれる構造的な情報には，各文が言及している話題や内容に関する情報と，文間の関係である接続関係に関する情報の２つがある．ここでは後者の，文間の修辞的な関係に着目し，「思考の流れ」としての文脈構造を考察する．

まず，文脈の捉え方について考察し，文脈構造の表現を形式化する．次にテキストから文脈構造を自動抽出する方法を定式化し，その方法を実際の論説文データに対して適用することによってその有効性を検証する．結論として，テキストが言及している対象分野の知識なしでも，文脈構造を良好に抽出できる第一ステップとしての見通しが得られた．

An Analysis of Rhetorical Structure**

Kenji Ono　Teruhiko Ukita　Sin'ya Amano

R&D Centre, Toshiba Corp.
Komukai-Toshiba-Cho 1,Saiwai-ku, Kawasaki, Japan 210,
tel.(044)549-2240 Japan

In this paper we discuss the global structure of Japanese articles, especially the global organization of Japanese dissertations. A text contains two kind of structual information: information about the subjects of the text, and information about the relations between each part of the text. We focus our attention to the latter, which appears as the rhetorical expressions in the text, and consider the text structure in the light of flow of thinking.

First, we discuss how to express the rhetorical structure of text and define the expression. Next, we devise a method of determining the rhetorical structure by analysing the configuration of the rhetorical expressions in texts. Finally, we evaluate the method by applying it to 2 Japanese articles, and show that, though the unique determination of the structure is beyond the scope, it can efficiently restrict the allowable rhetorical structures without the knowledge about the subject of the text.

＊ 本研究は，ＩＣＯＴからの委託により第５世代コンピュータプロジェクトの一環として行っている．

＊＊ This work is supported by ICOT (Institute for New Generation Computer Technology).

1. はじめに

本稿では日本語テキスト，特に論説文を対象にして，要約機能などの”文脈処理”の実現のために必要となる文章の全体の構造を分析する．文章全体に含まれる情報には，文間の関連を示す接続関係を中心として，話題の流れなどの要素が含まれる．ここでは，文間の修辞的な関係を文脈構造と捉え，「思考の流れ」としての文脈を考察する．

本稿では，まず文脈の扱いについて考察した後に，３節では文脈構造の表現方法を，４節ではその表現例を示す．その後，５節では文脈構造をテキストから抽出する方法を検討し，６節では５節で述べた構造抽出処理の具体例，およびそれを２文書（約１５０文）に適用して得られた評価結果とを記す．

2. 文章の構造

文章の構造という場合，書式的なものとして，章や節の構造，パラグラフなどが，思い浮かぶ．また逆に内容的なものでは，起承転結，序破急などの”修辞法”的な言葉を思い浮かべることが多い．本稿では，書式的な構造ではなく，それ以下の部分，即ち，節に分けられている文書では，節の内部の文章の構造を考える．

文脈構造は，文と文，さらにそれらが結合されたものの間にある接続関係を記述したものである．

テキスト中の修辞的関係に基づいた文脈構造の表現方法については，国文法学者による包括的な研究がなされている（永野[86]等）．その方法は大別して

（ａ）接続詞等に基づいたもの
　（テキストの構成要素間の相対的関係に基づいたもの）
（ｂ）名辞，代名詞等の現れ方に基づいたもの
　（主語，主題に基づいたもの）
（ｃ）述語の現れ方に基づいたもの
（ｄ）陳述の仕方に基づいたもの
　（文のタイプに基づいたもの）

に分類できる．従来の計算言語学においては（ｂ），（ｃ）によるものが主流であった．

辻井[88]は社説を題材として文を６つのタイプにわけ，それらの間にｃｆｇ（文脈自由）規則を設けて文章の構造化を考察しているがこれは（ｄ）に部類する．ここで考えられている構造は，各文の内容の間の関係（事実と判断，原因と結果，等）及びそれら内容と話者との関係に基づいたものであり，また文章全体の構造ではない．

（ａ）の立場による文脈の表現の研究としてはMannら[88]，所[86]等がある．Mannは文脈構造の最小単位をclauseとし，構造間の関係としては約25個定義している（solutionfood，purpose，circumstance，joint，evidence，elaboration 等）．しかしそこでは，例示，根拠といった修辞的関係と因果関係，目的ー手段関係といった非修辞的関係が混同されており，また表層表現からその構造を作る方法が述べられていない．

所は論説文の構造化にあたって，文を単位とし，構造間の関係を３つのレベルに分けて考察している．また表層的特徴とそれら構造間の関係との対応についても述べている．しかし構造化は人間による読解を前提としており，計算機で直接実現できるものではない．

本論文では所[86]を踏まえて（ａ）の立場で論説文の構造化を考察する．

論説文に於いては，その論旨の展開に構造的なものを感じることができる．これは論者が自らの言説を相手に判らせるため，言説を細分化し，因果関係などでそれらを関係付けてまとめあげ，次にそれらを特定の修辞方法によって並べていくという経緯をへて論説文がかかれているからである．こうしてつくられた文章の構造は人間の思考の一般的な流れに沿ったものとなっており，その文章は理解し易いものとなる．こうした思考の流れとして文脈構造をとらえ，その構造を表現しようとするとき，以下の２つの点を前提とすることは妥当であると思われる．

（ａ）構造はツリー状に表現することができる．

論説文が細分化された個々の言明から構成されていくときに用いられているものは，原因と結果，主張と例示，前提と結論，一般論と各論といったような２項関係である．これらの関係が再帰的に用いられることによって論説文は構成されている．従ってこの構造はツリーで表現することができる．

（ｂ）構造の最小単位を文とする．

ここで構造化しようとしているのは書き手の思考の流れである．同内容の事柄を表現する際，書き手によって文の切り方に相違があるかもしれない．しかし，各々の書き手が１文を一つのまとまりと感じていることは同じである．むしろ論旨の展開するうえで扱い易いような単位として思考をまとめたものが，その書き手にとっての１文といえる．

従って，思考の流れとしての構造を考えるときに，文（書き手にとっての文）をそのまま単位として用いることは自然である．同様に，パラグラフや段落分けがある場合には，そのことを構造化の際に積極的に利用する．

もう一つの理由は，思考の「流れ」とその「内容」には直接の対応関係はないという見方によるものである．これは，修辞的表現と照応，参照表現や名辞の反復といった現象とは基本的に独立であるという観察に基づいている．１文中に現れる個々の述語や名辞はその文の内容と関係するものであって，思考の流れ，即ち他の文との相対的関係とは無関係である．従って，１文中の個々の述語や名辞に立ち入ることは不必要と考えられる．

3. 文の修辞方法

論説文に於ては，文を連ねて文章を展開する際の修辞方法は幾つかの類型で押えることができる．その修辞方法は，書き手の思考の流れと対応するものである．そして作られた文章はそこでもちいられた修辞方法によって構造化されたものとなっている．また，その類型は接続詞等の表層表現と密接な関係をもっている．従ってそれらの表層表現から逆に元の修辞構造をある程度復元することができる．

その修辞方法には大きく２つのレベルがある．論旨を展開するうえでの最小単位となる１つの言明を構成する際に用いられるものと，それらの言明を関係づけていって論旨を展開する際に用いられるものとである．文を単位として見た場合，どちらの修辞方法も文あるいは文のまとまりの間を「関係」づける表現として現れてくる．前者を言明のレベルの関係，後者を思考のレベルの関係と以降呼ぶことにする．各レベルは文脈構造を構成する際の機能によって特徴づけられ，また文脈における諸現象（参照表現，話題表現）の現れ方によって特徴づけることができる．しかし，その２つのレベルの境界が不明瞭である場合もある．

図３．１にその該略を示す．

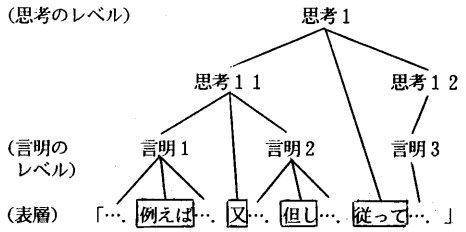


図３．１　言明，思考のレベルとその構造

本稿では国文法における接続詞（鈴木[73]）および複数のジャンル（演説文，経済記事，科学解説記事）の論説文数篇（約１０００文）から抽出した接続詞相当語句約８００を分類し，現在言明のレベルの関係として４種類，思考のレベルの関係として９種類定義した．

（ａ）言明のレベルの関係及びその修辞方法

特徴
- ２～３文のスパンで成立する．
- 次のレベル（思考のレベル，論旨を展開するレベル）での単位となる．
- １言明中に話題は１つのみ存在する．
- 一つの話者の言明（判断，感情，認識）を表現したもの．
- 中心となる部分が１つだけあり，他の部分はそれに付属して，中心となる部分の内容を明瞭にしたり，補足したりしている．

		記号
①例示	例：" 例えば"	▽
②重複	例：" というのは"	≡
③理由	例：" なぜなら"	←
④補足，婉曲	例：" 但し"	～

（ｂ）思考のレベルの関係およびその修辞方法

特徴
- 言明を単位とし，それらの間を関係づけていくもの．関係づけられた言明の間をまた関係づけていくときにも用いられる．パラグラフや段落の間の関係もこれでとらえることができる
- 展開される論旨の構造自体の表現となっている．

以降この関係でつながれた言明全体を「思考」とよぶことにする．

			記号
並列型：	①並列	例：" また"	＋
	②対比	例：" 一方"	－
直列型：	③順接	例：" 従って"	→
	④逆接	例：" しかし"	×
同列型：	⑤同列	例：" すなわち"	＝
指示型：	⑥指示	例：" 以下に…述べる"	※
	⑦参照	例：" 図２に…示す"	※
転換	⑧転換	例：" さて"	／
概括	⑨概括	例：" 結局"	□

4．　文脈構造の表現

前節で分類した修辞方法は，すべて２つの文あるいは文のまとまりの間を「関係」づけるものである．そして，どの文がどの文と，また文章のどの部分がどの部分とどう関係づけられているか，を記述したものが文脈構造である．

以下，ある文章の文脈構造の表現例を示す．

「Ⅰ－①火力発電設備は大幅な負荷調整能力と運用の多様化の要求にこたえなければならないという宿命を帯びている．②このような過酷な条件下でもなお長期的な高信頼度運用を達成するためには，まず，火力発電設備を構成する各機器の安全性の確保が前提となる．
Ⅱ－③一方，運転状態監視の強化や計画的予防保全などにより，定期点検間隔の延長化や，補修業務の合理化による各機器の長寿命化が試みられている．④このためには，運転中の機器の異常を早期に検出し，原因の分析や，対策に結び付く適切な情報の提供，あるいは，経年的な変化傾向を把握するなどの，きめ細かなデータの管理と分析，およびそれらの診断のための技術の確立が必要となる．
Ⅲ－⑤これらの判断は，これまで運転員あるいは補修員が監視計器などからの情報を，一般的判定基準に自己の経験を加味して行ってきたものであるが，⑥運用の多様化と介在するシステムの高度化とともに，判断を必要とする情報量も飛躍的に増加しつつあるのが現状である．
Ⅳ－⑦このような理由から，各種の情報を整理し加工して，運転や保全のための情報を提供する，計算機による診断システムの開発が急がれている．
Ⅴ－⑧以下に，火力発電設備の診断システムに関する概況およびその具体例を紹介する．」

（東芝レビュー1988.9月号 p710 「火力発電設備の診断システム」から抜粋）

上記の文章は次のように構造化される．

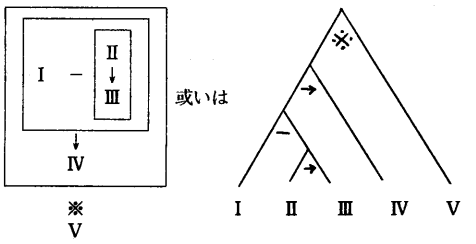
パラグラフ内：

Ⅰ	①→②
Ⅱ	③→④
Ⅲ	⑤×⑥
Ⅳ	⑦
Ⅴ	⑧

パラグラフ間：

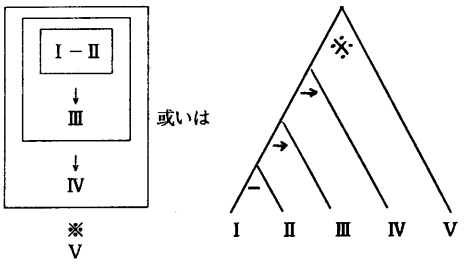
（（Ⅰ－（Ⅱ→Ⅲ））→Ⅳ）※Ⅴ

これは以下の表現と同等のものである．



尚，本例文については，パラグラフ間の構造として次のような構造も考え得る．

（（（Ⅰ－Ⅱ）→Ⅲ）→Ⅳ）※Ⅴ



このように，正解の構造候補が１つとは明確に限定できない場合もある．

5．　表層表現からの文脈構造の抽出

本節では，与えられた文章からその表層的修辞表現を手がかりとしてその文章の文脈構造を抽出する方法を考察する．現段階では３段階に分けて考えている．

5．1　単文化処理

２．で述べたように文を単位として文脈を捉えようとすると、書式的には文（２つの読点で囲まれた部分）でありながら論旨展開の機能で考えた場合には１文以上あるいは１文以下に相当するものが存在する．前者は文中に等位接続詞が使われているような複文であり，後者は１接続詞に還元できるような常套表現からなる単文である．例えば，

”・・・であるが，しかし・・・が問題である．”

といった文は，論旨展開のうえでは逆接の関係で関係づけられた２文である．よって，これらの文については

”・・・である．しかし・・・が問題である．”

のように２文として扱う．又，

”・・・．そればかりではない．・・・．”

のような文は機能的には「さらに」と同等であるので，

”・・・．さらに・・・”

と同等のものとして扱う．このようにして，まず文脈構造の単位となる文を決定する．

5．2　関係の取り出し

５．１の単文化処理が済んだ文章に対して，主に各単文の先頭の接続的表現をみて，それが３．で分類した修辞関係のどれに該当するかを決める．複数候補が考えられるときはすべての可能性を考慮する．この処理では，従来の国文法における接続詞のみならず論旨展開の際に用いられる修辞的表現一般を用いて関係を抽出している．それら接続的表現が無い部分については，現在デフォールトとして順接或いは並列の関係を代入している（便宜的にラベル´？´を付与する）．

パラグラフ間の関係の抽出は，各パラグラフの先頭の文の接続的表現から決定している．

5．3　構造化

５．２の処理で取り出した修辞関係の１パラグラフ内での並びを見て，そのパラグラフの修辞的構造を決定する．この修辞構造とは４節の例に示すような２分木である．

この処理は通常の構文解析処理に類似したものであるが，文法のようなrigid な規則は用いず，その代わりに，２つの前提と規則（優先規則）を用いて，可能な修辞構造の候補を順位づける処理を行っている．パラグラフ間の構造化もパラグラフ内の構造化と同じように行う．

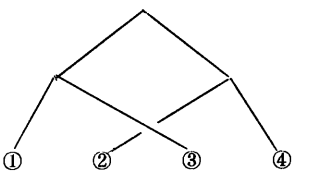
その前提と規則は以下の通りである．

前提１：　１パラグラフ（節，段落）に対して一つのツリーが構成される．

この前提は，各パラグラフあるいは各段落がそれぞれ思考の流れの上では単位となっており，独立して閉じた構造を作っているということを主張している．パラグラフや段落は書式的なものであり，内容の構造はそういった表層的構造とは必ずしも一致しないという見方もあるが，２節でも述べたように，「内容」ではなく「思考の流れ」として文脈を捉えるにあたっては書式的情報を直接利用することは自然であろう．

前提２：　ツリーは非交差である．

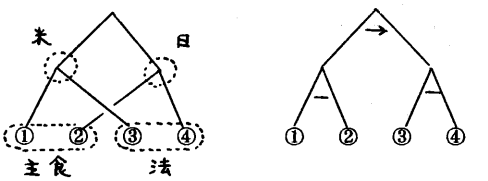
これは次のような構造は認めないということである．



この理由として，以下のような文章を考えてみる．

”①米国では主食はパンだが②日本では白飯である．従って③米国の法的規制は小麦粉を重点的にあつかっているが④日本では米が主対象である．”

という文章の構造を考察してみる．この文章の構造として以下の２つの構造が直観的に想定される．



ここで交差部分を含む左の構造は文章の「内容」あるいは「話題」の構造であり，右の構造は文章の「流れ」の構造である．論説文においてはこのように内容的に前後する文章はあっても，思考の流れが前後するものがあるとは考えにくい．従って前提２は妥当であると思われる．

この前提を認めると，可能な構造の総数は与えられた文章に含まれる文の数のみによって次のように決定される．

文数	候補数
1	1
2	1
3	2
4	5
5	1 4
6	4 2
7	1 3 2
8	4 2 9

次に，各修辞関係の相対的関係を考察し，そこから導かれる修辞関係の並び方に関する規則について述べる．

起承転結型の代表的な文としてよく引用される文章に以下のものがある．

「 紀州堺の糸屋の娘 …①
姉は１８妹は１７ …②
戦国武将は弓矢で殺す …③
糸屋の娘は目で殺す」 …④

この文章の構造は以下のようになると思われる．

（1→2）→（3－4）

このような文章構造は論説文にはあまり見当らないことが指摘されている（所[86]）．これは論説文の書き手に修辞能力がないからではなく，「転」の部分に含まれる，読み手に論旨展開を見失わせる効果が論説文のレトリックとしては不適当であるからである．これは構造の上ではアンダーラインを施した部分に相当する．

この部分が思考の流れにそぐわない理由は，①の部分は②の部分へと順接し，③の部分は④の部分と対比の関係にありそれぞれは思考の流れに即しているのだが，②と③とは無関係であるからである．③は④との関係においてのみ存在しているのだが，表現上④の前に位置せざるを得ず，直接には無関係な②と並ばざるをえなくなっているからである．

これを踏まえて，この様な構造を含むものが論説文の構造の候補として挙がった場合には，その構造を含まない別の候補を優先するという規則を考えることが出来る．この考えを以下のように各関係間に敷えんし，以降規則１と呼ぶことにする．

規則１：以下の構造を含むものは，棄却する．

[…，直列型関係 （ X 直列型関係 …]
[…，同列型関係 （ X 直列型関係 …]
[…，指示型関係 （ X 直列型関係 …]
[…，指示型関係 （ X 並列型関係 …]

例えば，

① だから ② だから ③
順接（→） 順接（→）
直列型 直列型

という修辞関係の列に対しては

（①だから②）だから③
（① → ②） → ③

という構造は許されるが

①だから（②だから③）
① → （② → ③）

という構造は棄却される．

規則１中の´X´は，その部分には１つの「言明」あるいは「思考」を構成する文あるいは文のまとまり全体が入ることを示している．形式的には何かの部分木（のルートノード）がそこに入ることを示している．従って，

…→（文→…
…→（（文＋文）→…
…→（（文＋（文▽文））→…

といった構造は全て不適合とされる（ここで´→´は「順接」，´▽´は「例示」，´＋´は「並列」の関係記号）．

同様，規則１に挙げられたものほどではないにしろ，やはり論説文の構造としてあまり好ましくないと思われる構造として以下のものがある．これらをまとめて規則２とする．

規則２：以下の構造を含むものは，優先度を下げる．

[…，直列型関係 （ X 並列型関係 …]
[…，並列型関係 （ X 直列型関係 …]
[…，同列型関係 （ X 並列型関係 …]

例えば

① さらに ② だから ③
並列（＋） 順接（→）
並列型 直列型

という修辞関係の列に対しては

①さらに（②だから③）
① ＋ （② → ③）

という構造より

（①さらに②）だから③
（① ＋ ②） → ③

という構造が優先される．

次に

…→（A→… （a）

という構造と

…→（（A→… や …→（（（A→… （b）
1 2 1 2

といったような構造を比較してみる．思考の流れという見地からはこれらの間に共通性を見いだすことが出来る．すなわち，アンダーライン２を施したような位置にくる修辞関係（本例では順接）とその表層的な先行部分であるアンダーライン１の位置にくる修辞関係（本例では順接）とは，構造上のレベルが違

うから全く無関係という訳にはいかず，表層的に連続であるということから何等かの影響を及ぼしている．

よって，（b）のような場合にまで拡張して　規則1，規則2を使うことを考え，それらをそれぞれ規則1’，規則2’とする．

規則1´：以下の構造を含むものは，棄却する．

[…，直列型関係　(…（　X　直列型関係　…]
[…，同列型関係　(…（　X　直列型関係　…]
[…，指示型関係　(…（　X　直列型関係　…]
[…，指示型関係　(…（　X　並列型関係　…]

規則2´：以下の構造を含むものは，優先度を下げる．

[…，直列型関係　(…（　X　並列型関係　…]
[…，並列型関係　(…（　X　直列型関係　…]
[…，同列型関係　(…（　X　並列型関係　…]

現在は規則1と1´の間に（2と2´の間に）有意義な区別理由をみつけておらず，一緒のものとして扱っている．

これら以外にも，個別的な規則，個々の修辞的表現に対応した細則が多く考えられる．まず規則1，1’2，2’に着目して6節でその規則の妥当性を検討する．

6．　分析例

2種類のデータに関する分析の例を示す．

［分析例1］

3章で述べた方法で，4節の文章がどの様に構造化されていくかを示す．
単文化処理及び関係抽出処理は以下のようになる．

" ①…帯びている．②このような…前提となる．"

順接

" ③一方…試みられている．④このためには…必要となる．"
対比　順接

" ⑤これらの…ものである｜⑥が，…現状である．"
順接　順接or逆接or対比

" ⑦このような理由から…急がれている．"
順接

" ⑧以下に…紹介する．"
参照

これらの文章は5つのパラグラフに分かれており，パラグラフ単位に考えると，仮定1，2を満たす構造は14個考えられる．その優先順位は図6．1のようになる．
4節で記した正解と考えられる構造は9と14に相当する．

	構造候補	規則 1	1’	2	2’	
1	[1－[2→[3→[4※5]]]]	1	0	1	0	
2	[1－[2→[[3→4]※5]]]	0	1	1	0	
3	[1－[[2→[3→4]]※5]]	1	0	0	1	
4	[[1－[2→[3→4]]]※5]	1	0	1	0	
5	[1－[[2→3]→[4※5]]]	0	0	1	1	
6	[1－[[[2→3]→4]※5]]	0	0	0	1	
7	[[1－[[2→3]→4]]※5]	0	0	1	1	
8	[[1－[2→3]]→[4※5]]	0	0	1	0	
9	[[[1－[2→3]]→4]※5]	0	0	1	0	正
10	[[1－2]→[3→[4※5]]]	1	0	0	0	
11	[[1－2]→[[3→4]※5]]	0	1	0	0	
12	[[[1－2]→[3→4]]※5]	1	0	0	0	
13	[[[1－2]→3]→[4※5]]	0	0	0	0	
14	[[[[1－2]→3]→4]※5]	0	0	0	0	正

図6．1　各構造の優先順位

各規則の欄で1となっているところが，その構造候補がその規則に抵触していることを示している．
この表から，

① 規則1及びその拡張である規則1’を用いると構造候補は7つにまで絞られ，正解と思われる構造はどちらもそのなかに残っている
② しかし，より弱いプリファレンス規則である規則2及びその拡張である規則2’を用いてさらに絞ると，候補は2つにまで絞れるが正解候補の1つが棄却されてしまう

ことが判る．

［分析例2］

「　　2.2　ソフトウェア資産の活用の幕あけ

①ＩＢＭ　ＰＣ／ＡＴ用に開発された数万種類のごく一部の欧米のソフトウェアがわが国のパソコン用に日本語化され、大いに活用されている。②しかし、これらは氷山の一角であり欧米にはわが国でまだ知られていない、業務，業種用の優れた膨大な量のソフトウェアパッケージがある。③例えば、農業用ソフトウェアパッケージを例にとると、欧米には優に１００本を超えるソフトウェアがある。④その内訳をみると、最適出荷時期を決める市場価格プロット，散布，点滴，噴水など最適な水のまき方を決める潅漑（かんがい）ソフトウェア，連作障害を防ぐための耕地と作物の履歴管理，地形と地質からみた作付コンサルタント，乳牛飼料管理，種まきスケジュールなど、また、利益予測，農機具の減価償却などの経営支援と非常に多種多様でかつユニークなソフトウェアがそろっており、情報産業が農業を全面的に支援している。⑤一方、わが国で開発された農業用ソフトウェアパッケージはごくわずかでしかない。⑥これは欧米においてはパソコンがそのソフトウェアの豊富さによってあらゆる産業、あるいは個人生活まで普及している例である。⑦最近わが国で生まれたソフトウェアパッケージの中から欧米に輸出されるものも少しずつではあるが出てきている。⑧世界標準アーキテクチャがわが国でも普及するようにしていくことこそ、世界のソフトウェア資産の活用の幕あけになるものと確信する。」

（東芝レビュー1988.6月号p 490
「パーソナルワークステーションと新しい潮流」より）

単文化処理及び関係抽出処理をした後，この文章は以下のように表現される．

[①－or×②▽③→④－⑤→⑥？⑦？⑧]

この７つの関係記号の列に対し可能な構造は４２９個考えられる．それらに対し各規則を適用すると，以下のように候補が絞られてくる．

規則１のみ適用　…３８７個

規則１と１´を適用　…３６３個

規則１と２を適用　…６８個

規則１，１´，２，２´を適用　…２８個

正しい構造は次のようなものであると考えられる．

（（（①－②）▽（（（③→④）－⑤）→⑥））＋⑦）→⑧
（（（①×②）▽（（（③→④）－⑤）→⑥））＋⑦）→⑧
（（①－（②▽（（（③→④）－⑤）→⑥）））＋⑦）→⑧
（（①×（②▽（（（③→④）－⑤）→⑥）））＋⑦）→⑧

これらの構造はすべて，最後の２８個のなかに含まれている．

次に，同様の分析方法によって論説文２文書全体について調べた結果を述べる．
解析結果の表現は，論説文中の構造候補が複数でてくる部分の各々について

１）可能な候補数
２）規則１＋１’で絞った候補数，及び
　その中に正解候補が含まれているかどうかの判定
３）規則１＋１’＋２＋２’で絞った候補数，及び
　その中に正解候補が含まれているかどうかの判定

を列記したものとなっている．尚，正解候補が１つに決められないものについては，最大５個として複数認めている．
判定は以下のように表記する．

○…正解候補がすべて含まれている
△…一部棄却されている
×…すべて棄却されている

文書１：東芝レビュー1988.9月号 pp. 710-714,
「火力発電設備の診断システム」

６章２０段落
４８パラグラフ
８８文

このうち構造候補が複数出てくるパラグラフ，段落はそれぞれ１６個，８個の計２４個である．分析結果を図６．２に示す．

章，節	文数	１）	２）	３）
（括弧内はパラグラフ数）				
[1] パラグラフ間	(5)	14	7 ○	2 △
[2] 第２パラグラフ	3	2	2 ○	2 ○
第４パラグラフ	3	2	2 ○	2 ○
第５パラグラフ	5	14	14 ○	5 ○
[2] パラグラフ間	(6)	42	4 ○	4 ○
[3] 第１パラグラフ	4	5	5 ○	5 ○
[3.1] 第１パラグラフ	4	5	4 ○	1 ○
第２パラグラフ	3	2	2 ○	1 ○
第３パラグラフ	5	14	9 ○	9 ○
[3.1] パラグラフ間	(4)	5	5 ○	1 ○
[3.2a]パラグラフ間	(3)	2	2 ○	1 ○
[3.2b]第１パラグラフ	3	2	2 ○	2 ○
[3.2b]パラグラフ間	(3)	2	1 ○	1 ○
[4] 第１パラグラフ	3	2	2 ○	2 ○
[4.1] 第４パラグラフ	4	5	5 ○	2 ×
[4.1] パラグラフ間	(4)	5	5 ○	1 ○
[4.2] 第３パラグラフ	3	2	2 ○	1 ○
[4.2] パラグラフ間	(3)	2	2 ○	2 ○
[4.3a]第２パラグラフ	3	2	2 ○	2 ○
[5] 第１パラグラフ	4	5	5 ○	5 ○
[5.1] 第１パラグラフ	5	14	7 ○	1 △
第３パラグラフ	5	14	14 ○	14 ○
[5.1] パラグラフ間	(3)	2	2 ○	1 ○
[6] 第２パラグラフ	3	2	2 ○	2 ○

図６．２　文書１の解析結果

文書２：東芝レビュー1988.6月号pp. 490-492，
「パーソナルワークステーションと新しい潮流」

８段落
１２パラグラフ
７３文

章，節	文数	１）	２）	３）
[1]	3	2	1 ○	1 ○
[2.1]	8	429	384 ○	10 ×
[2.2]	8	429	363 ○	28 ○
[3.1]	10	4862	4862○	264 ×
[3.2] 第１パラグラフ	18	129644790	解析不能	
第２パラグラフ	9	1430	1430○	1430○
[4]	5	14	5 ×	2 ×
[5]	5	14	14 ○	5 ○
[6]	6	42	42 ○	42 ○
[7]	8	429	429 ○	132 ○

図６．３　文書２の解析結果

これらの結果から以下の結論を得た．
まず，各規則の妥当性（正解構造を棄却していないか）については，概ね良好な結果を得ることができた（特に規則1+1'について）．誤判定したものを分析した結果，その原因は連体詞，代名詞の扱いによるものであることが判った．
関係抽出処理では現在，連体詞或いは代名詞で始まり（”…．この…”，”…．それらは…”等）他に接続的表現が無い箇所については直列型関係としている．しかしそれらは，他の直列型の関係表現（”従って”，”だから”等）と同じようには規則１や２に従っていなかった．

例えば

" ①一方②従って ③"

という表層部分を含むものはすべて規則2に従っていたが,

" ①一方②これらは③"

という表層部分を含むものについては, 一部規則2に違反して,

…①－ (②→③) …

という構造になっているものも存在した.

次に各規則の候補制約力 (効果) については, 1/1 ～10/429 と場所によっておおきなばらつきが在ることが判った. 即ち, 各場所での接続関係の並びに多分に依存していることが判った. 平均すると, 規則1＋1´には約2／3, 規則2と2´まで用いた場合には約1／5にまで候補を絞る能力がある.

この値はまた文書の文体にも大きく依存している. 文書2は文書1に比べて1文が短く, 1パラグラフ中の文数は多い. このように文数が多いと接続関係の相対的組み合せが増え, 各規則が効き易くなり, 効果は大きくなる.

候補が殆ど絞れていないものについて調べてみると, それらは接続表現が乏しい部分であった. 5. 2で述べたように, 関係抽出処理では現在, 接続表現が無い箇所は, 並列型或いは直列型の関係として扱っている. しかしそのような箇所が多いと, 各規則があまり効かなくなってしまう. 文書2ではこれが原因となって候補が絞れていないケースが大半であった.

7. むすび

本稿では, 日本語テキストの文脈構造の分析を行った. 文章の構成を表現する文脈構造は, 文間の (文章の構成部分間の) 関係を記述する必要があるとの立場から, 文間の接続表現を抽出し, それにより, 2分木の形式で関係を記述する方法をとった. 具体的には, 約800個の接続表現を13種の接続関係に分類し, それら接続関係の組み合せについて考察し, 文章の構造として許されるものを規則化した.

これらの方式を実地に検査するために, 日本語の論文2編, 約150文を実際に分析し, 正しい形態素解析を想定して, 自動処理の可能性を検討した. その結果, 曖昧さは残るものの, テキストが言及している対象分野の知識無しでも, 接続関係から可能な構造を効果的に削減できることを示した. 今後, より多くのデータに関して検証を行っていき, 各種の規則の精密化とアルゴリズムの更新を進めていく.

[参考文献]

辻井 潤一：論説文における文脈構造, 日本学術振興会 文字言語・音声言語の知能的処理第152委員会第7回研究会資料 7－1, 1988.

所 一哉：現代文レトリック読解法, 匠出版, 1986.

永野 賢 ：文章論総説－－文法論的考察－－, 朝倉書店, 1986.

鈴木 一彦, 林 巨樹編集：品詞別日本文法講座6 接続詞・感動詞, 明治書院, 1973.

Mann,W. and Thompson,S：Rhetorical Structure Theory: A Framework for the Analysis of Texts. USC／Information Science Institute Research Report, RR-87-190, 1987.