ICOT Technical Report: TR-501

TR-501

日本語テキスト理解における文脈構造抽出法

木下　聡、小野　諭司、
浮田　輝彦、大野　真家（東芝総合研究所）

September, 1989

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03) 456-3191~5
Telex ICOT J32964

Institute for New Generation Computer Technology

日本語テキスト理解における文脈構造抽出法*

木下 聡　小野 顕司　浮田 輝彦　天野 真家

(株) 東芝　総合研究所

本稿では，テキスト理解と文脈構造との関係について整理し，テキスト理解のモデルを示す．そして，計算機により文脈構造を抽出するため構造抽出法について検討する．その抽出法では，文間の関係である接続関係を利用した思考上の制約に加え，話題表現や文のタイプなどの言語的情報を用いて構造の適切さや他の構造に対する優先性を調べて，優先度の高い構造を選び出す．さらに，構造化の実験を行って，言語的情報の利用により，接続関係のみを用いた場合に比べ，さらに構造を絞り込むことができることが確認できた．

Discourse Structure Extraction for Text Understanding**

Satoshi Kinoshita　Kenji Ono　Teruhiko Ukita　Shin'ya Amano

R&D Centre, Toshiba Corp.

In this paper we discuss the function of discourse structure in text understanding and present a model for its understanding. Next we devise a method for extracting the discourse structures of Japanese argumentative articles. The method utilizes two kind of information. One is the constraint over the flow of thinking, which is described as prohibition on preference of structures by using the rhetorical expressions in texts. The other is the linguistic clue such as topic, type of sentence and so on. Finally, we evaluate the method by applying it to 2 Japanese articles, and show that it can restrict the allowable discourse structures more efficiently than the method we presented before, 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

近年，会話文の理解やテキストの理解など広い意味での談話理解に関する研究が多数行われている．当初は，入力文中の代名詞が何を指しているかを判断する照応問題が主として取り上げられてきたが，このところ，会話やテキストそのものの構造を捕らえる試みが行われつつある．

会話においては，会話の相手と交互に発話を繰り返すことで形成される会話全体の構造を考えることができるだけでなく，個々の発話に関してもその構造を考えることができる．我々がこれまで行ってきた会話文理解の研究においては，会話の対象分野の知識を用いて発話の内容を理解するという観点から研究を進めてきたが［木下88, Ukita88］，一度の発話として処理すべき文は１ないし２文程度と仮定していた．しかし，会話の相手に対して伝えようと思うことがらが複雑であったり相手に予備知識がなかったりすると，背景や用語の説明をしたり，前の文を言い換えたりして相手の理解を助けることが必要な場合がある．その結果，発話を構成する文の数は増加するが，それらは単なる文の羅列ではなく，文間の関係によって組み上げられた１つのまとまりのある構造である．このようなことから，テキストの理解のみならず会話文の理解においても，各文の役割や文間の関係を認識し，その構造を取り出すことが重要となる．

我々は先に［小野89］で報告したように，その第１ステップとして，文間の修辞的関係を基に論説文の構造化を試みたが，テキストの構造を全て一意に決定することはできなかった．そこで本稿は，まず，テキストの理解と文脈構造の関係について考察し，テキスト理解のモデルを検討する．さらに構造化を行う際の手掛かりとして，先の構造化で用いた手掛かりに加え，文の述語や文末の助詞の表現などの言語的手掛かりを用いて構造を抽出する方法を検討する．

2　テキスト理解のモデル

2．1　テキストの生成と理解

（1）テキスト生成における文脈構造

テキストは，書き手が読み手に対して情報を伝達する１形態であることは明らかである．したがって，テキストを記述する書き手の心の中には，伝達したい情報の知識構造がまず存在すると考えられる．伝達したい内容は，１つの命題で表現できる場合もあろうが，ここでは１つのまとまったテキストで表現されるような，種々の事象のみならず，書き手の認識や態度を含んだ構造であると考える．この構造自体は，何を述べるかが決まった段階のものであり，それらをどういった順序で述べるかといった点で，順序付けられているものではない．

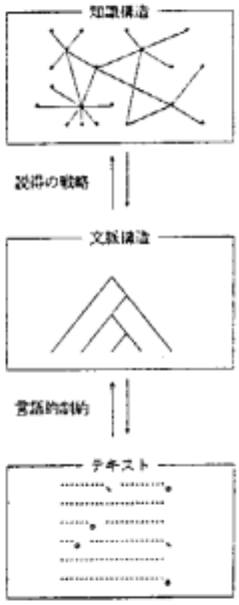

図２．１　文脈構造の位置付け

図２．１に知識構造からテキストが生成される過程を示す．まず，上で述べたように書き手が読み手に対して伝達したい情報の心的表象である知識構造がある．この知識構造は，言語表現であるテキストの形に表現するために，まずどのような内容から始めて，どのような順序でどう関係付けて記述すれば，読み手にとって分かりやすいかといった観点から別の構造に再構成されると考えられる．このようにして得られる構造が，我々のいう「文脈構造」である．文脈構造では表現の単位である文間の関係や順序が明らかになっている．そして，それらの文を最終的な言語表現に落とす際には，例えばLeech の４原則（処理可能性，明晰性，経済性，表現性［Leech83 ］）のもとで，種々の言語的制約に基づいてテキストの形になると考えられる．すなわち，文間の関係を読み手が認識するのが難しいと思われる場合には，接続詞を用いたり，簡潔さのために代名詞を用いたり語を省略したりするわけである．

ところで，書き手の持つ知識構造は，始めから決定されていて変化しないというわけではない．例えば，文を書いている際に，書きたいことがらが徐々に変化することがある．その場合には，当然のことながら文脈構造も変化することになるが，できあがったテキストを読む読み手から見れば，始めからある知識構造があって，それをもとにテキストが生成されたと考えてかまわない．また，伝達すべき中心的情報が与えられたときに，それから読み手のレベルなどに応じて，どの程度詳細に述べるかを決めて，それをカバーする知識構造を決定するプロセスがあると考えられるが，ここではその過程までは考えない．

（2）テキスト理解の過程

テキスト生成との対応で考えると，テキストの理解は，生成と逆方向の過程であり，与えられたテキストから

書き手が伝達したいと考えている知識構造を復元することである．そしてその１ステップとしてテキストから文脈構造を得る過程があると考えられる．特に論説文などでは，文脈構造を求める際に基本となる手掛かりは，テキスト中に現れる接続詞などの言語表現（以下「接続表現」と呼ぶ）である．読み手はこれらを手掛かりして，例示や理由といった文間の修辞的関係を認識し構造化を行っていく．

これまでのテキスト理解を振り返ると，まず，初期のアプローチの１つとして非言語的知識を利用したアプローチがある．これは，Schank らがスクリプトと呼ばれる典型的な事象のシーケンスに関する構造的知識をもとにして，一連の入力文で表現された事象同士を関係づけることでテキストを１つのまとまりのある構造として捕らえようとしたものである［Schank77］．このアプローチは，知識に対しての依存度が大きく柔軟性に欠けるという問題点があり，その後知識の動的な構成の試みが行われている．Johnson-Laird は，談話の構造を捉えるものとして，メンタルモデルを挙げている．彼の議論によれば，与えられた文章がでたらめな文の羅列でないための必要十分条件は，それから単一のメンタルモデルが作られることであるとしている［Johnson-Laird83］．そして，そのための要因として，同一指示（co-reference）と一貫性（consistency ）を挙げている．しかし，そのメンタルモデルの具体的な表現方法などは述べられていない．どちらの部分も入力文は事象の並びであり，それらの間に結束性を見出すことに主眼がおかれており，論説文などのように単なる事象のみならず，書き手の認識や意見が述べられていたり，図表への参照の指示が含まれているテキストを処理するモデルとしては不適当である．

心理学においてテキストの構造の記述を試みたものとして，物語文法（Story Grammar ）がある［Rumelhart75］．この理論は，民話などの物語に現れる定型的なパターンを，文脈自由文法などを用いて記述することで，物語全体を１つの構造として認識しようとしたものである．しかしながら，記述される文法規則には，構文規則において存在するような心理的実在性はほとんどないし，構文上の規則のように厳密な規則性が存在するわけではないので，全ての物語を認識できるような一般的な文法を記述することは実質的に不可能である．また，文の内容を分析し，それを「場面設定」であるとか「挿話」であると判定する明確な手段が欠けているといった問題がある．

また，言語学におけるテキストの研究として永野は，テキストを文間の連接関係，連鎖関係，統括関係の３点から分析し，それぞれの次元におけるテキストの構造化を行っている［永野86］．また，所は，文章の構造化の際の修辞的関係として，陳述のレベル及び思考のレベルの修辞的関係，そして思考のレベルによって構成される叙述型の存在を示し，これらの３つのレベルでの構造化を提案している［所86］．しかし，これらの研究においては，構造の抽出は人間による読解を前提としており，計算機による処理に対する考察は行われていない．

Mannらは，文脈構造を作る単位を節（clause）とし，それらの間の関係を２５個程度に分類して，テキストの構造を記述する試みを行っている［Mann87］．しかし，入力文からの構造の組立に関する議論はなされていない．

辻井は，論説文の構造を言語的知識のみを用いてテキストの構造を抽出する試みを行っている．そこでは，文のタイプを事実文，判断文，要望文などに分類し，それらのタイプ間の可能な連続を文脈自由文法で記述しておくことで，入力テキストから修辞的構造を抽出している［辻井88］．また，Schaらは，文脈の構造に関する各種の知見をまとめ，やはり，テキストの構造を文脈自由文法で表現し，それを用いて構造化する枠組みを提案している［Scha88］．しかし，物語文法での場合と同様，文章の多様性を考えた場合，全てのテキストの構造を表現するだけの文法規則を予め用意することは不可能であり，実際的な方法ではない．

Cohen は，接続詞などの言語的手がかりと，文が表現する命題間の関係をもとにテキストの構造化を行うモデルを提案しているが，計算機上での実現には至っていない［Cohen87］．

我々は，接続詞などの修辞的関係の表現を手掛かりとして文脈構造の抽出を行っている［小野89］．しかし，このような接続表現が全ての文間に現れるわけではない．さらに接続表現が明示されているからといって，構造が一意に決定できるわけではないという問題が残されている．本稿では，計算機によるテキスト理解のモデルについて整理し，接続表現以外の言語的手掛かりまで含めた文脈構造の抽出法を検討し，その有効性を探る．

なお，このようにして得られる文脈構造は，テキストの要約などに用いることができるほか，代名詞などの照応表現の指示先を解析する際に利用することもできる［例えば辻井88］．

２．２　計算機によるテキスト理解のモデル

図２．２にテキスト理解のモデルの概略を示す．２．１で示したようにテキスト理解の最終目標は知識構造を得ることであり，そこに至る１ステップとして文脈構造を考える．テキストから文脈構造を求める際に基本となる手掛かりは，テキスト中に現れる接続詞などの言語表現（以下「接続表現」と呼ぶ）である．読み手はこれら

を手掛かりして，例示や理由といった文間の修辞的関係を認識し構造化を行っていく．しかし，接続表現が明示されているからといって，構造が一意に決定できるわけではない．なぜなら，文と文との関係は連続する２つの文の間の関係とは限らないからである．例えば，次のような文の並びがあるとする．

①・・・。
②例えば，・・・。
③又，・・・。

構造の表現については，次章で詳しく述べるが，上の文の並びに関して次の２つの構造化が可能である．

（1）　（（①　例示　②）　並列　③）
（2）　（①　例示　（②　並列　③））

これは，（1）では文③は文①で述べられていることと並列の関係にあるのに対し，（2）では②③２つが両方とも①の例を述べていることを表している．

このように，接続表現が明示されている場合でも，それだけでは構造を一意に決定することはできない．また，このような接続表現が全ての文間に現れるわけではない．読み手が容易に認識しうると考えられる関係は明示しないためである．その場合には，読み手は文の内容から，文間の関係を認識する必要がある．そのため，非言語的知識の利用は不可欠である．しかし，現段階においては，常識と呼べるような広範囲に渡って均質に表現された知識を計算機上で利用するには至っていない．したがって，現在のモデル化では，言語的知識の利用を中心としたモデルにならざるをえない．

そこで，本研究では次の２つをもとに構造の絞り込みを行う．

・（接続関係を用いた）思考上の制約
・接続表現以外の言語的手掛かり

以下では，それぞれについて説明する．

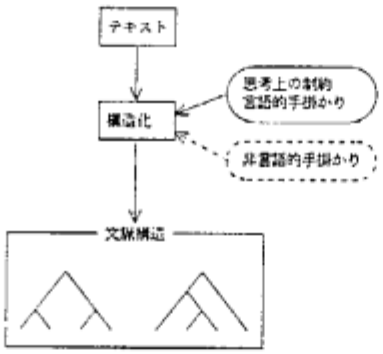

図2．2　文脈構造の抽出過程

表2．1　構造化の手掛かりの分類

項　目	使用回数
思考制約	10
話題表現	5
同語反復	8
類語反復	3
文末表現	3
文のタイプ	9
構造の類似性	4
代名詞，連体詞	4
同義性	8

（a）思考上の制約

テキストは，思考の単位となる文を修辞的な関係で組み合わせて構成されたものであり，特に論説文などでは，これらの関係が言語的な表現として明示される場合が多い．したがって，テキストから文脈構造を取り出すには，テキストに現れる修辞上の手掛かりを用いることが基本となる．

しかし，文間に修辞上の手掛かりがあるからといって文脈構造が一意に決定できるわけではない．その場合には，思考の方法に関する制約を利用することができる．先に述べたように，書き手の論旨は，思考の単位となる文を修辞的関係によって組み合わせていくことで形作られる．しかし，２つの思考単位の間をそれらの関係で結合しさえすれば，論旨の展開として適当であるかというと必ずしもそうではない．それらの中には，われわれの思考の認知的な能力の点からみて，不適当な構造が存在するからである．また，論旨が次々と展開される場合，そこに思考の流れといったものを感じることができるが，そのような場合によく使用される構造もあると思われる．このように思考の展開上，優先される構造やほとんど認識し得ない構造があるが，これらをまとめて「思考制約」と呼ぶ．

この思考制約は，絶対的なものではない．そのため，これらをもとにして構造を組み上げることは適切ではなく，制約として可能な構造を絞り込む際に用いるのが適切であると思われる．制約の具体例は次章で述べる．

（b）接続表現以外の言語的手掛かり

接続表現は，構造化において基本的情報の１つであるが，前にも述べたようにそれらだけでは構造を一意に決定することはできず，やはり文間のつながりを調べることが必要になる．表２．１は，調査した文献（科学技術論文３編の一部）において，パラグラフ内の文を人間が構造化する際に手掛かりとなった要因をまとめたものである．各項目の使用回数は，注目している文が他の文又はその集まりからなる構造と関係があると判断するのに主として使用したと考えられるもののみを数えており，２つ以上の項目が関係する場合でも，どれか１つを選んでいる．それらの中で，構造化の手掛かりとなっている言語的要素は，話題表現，文のタイプ，文末表現，同語反復などである．

・話題表現は，［浮田89］で述べたように，文が述べている対象を示すものであり，既に前方の文で現在対象となっている概念が提出されている場合には，その文と展開（事象で述べる修辞関係では「追加」と呼ぶ）や対比といった直接的な関係があるということができる．したがって，文脈構造を作る際に，これらの文間の関係が近い構造を優先する理由とすることができる．また，同語反復や類語の反復も話題表現に準じる形で構造化の手掛かりとして用いることができる（述語の反復も，ここでは同語反復としてカウントしてある）．

・文末表現は，テンスやアスペクトなどを含め，共通の性質を持つ文のグループ化に利用できる．例として，先に示した構造の曖昧性を持つ文の並びを考えてみよう．

①・・・～している。
②例えば，・・・～である。
③又，・・・～している。

①～③の文末表現が上記のような場合，次の構造である場合が多い．

（（①　例示　②）　並列　③）

・文のタイプは文を一般的な事実を述べている「事実文」であるとか，書き手の認識や意見を述べた「意見文」などに分類したものである．これは，パラグラフの先頭もしくは最後に意見文が現れた場合，その意見文が残りの事実文からなる部分構造を統括するような構造となる場合が多いことから，この形の構造を優先することができる．

もちろん，人間の場合にはすぐに非言語的知識を用いて文の内容を理解し，内容レベルで文間の関係を認識するため，表層的手掛かりはあまり意識することはないと思われる．しかし，ここでの分類では，表現上の手掛かりを第１とし，それらの手掛かりを見出しえない場合や判断の基準となりえない場合に，構造の類似性などの項目に分類してある．

同義性の典型例はいいかえである．いいかえの場合，「すなわち」といった接続表現がないと，文の内容を見て，それらが同じことがらを述べていることを認識する必要があるり，そのためには類似した概念をどの程度同一視するかといった推論が必要となる．また，代名詞や連体詞は，表現そのものは容易に検出しうるが，その照応先や連体詞のスコープが分らなければ構造化に利用することは難しく，それには知識を使った処理が必要となる．

以上のように，構造化の手掛かりとして知識を利用するものもあるが，思考上の制約と言語的手掛かりによる制約を用いることにより，ある程度構造の絞り込みを行って，テキストから文脈構造を抽出することができると考えられる．しかし，最終的には，非言語的知識を利用した処理が必要となる．

３．　文脈構造の表現

３．１　修辞的関係の分類

［小野89］で示したように，我々は［所86］をもとに，修辞関係を大きく２つのレベルに分けて考えている．１つは言明のレベル，もう１つは思考のレベルである．言明のレベルは，論旨を展開する上での単位となる１まとまりのことがら（これを「言明」と呼ぶ）を述べるためのレベルであり，また，思考のレベルは，それらの言明を関係づけていって論旨を展開するレベルである．図３．１にテキストの表層構造と，言明及び思考のレベルでの文のまとまりと展開の模式を示す．ただし，言明のレベルと思考のレベルの境目は必ずしも明確ではない．１文で１言明を表す場合もあるし，１パラグラフでありながら１言明に満たないと思われる場合もある．

以下に言明と思考のレベルの関係の性質と，修辞方法を述べる（ここで示す関係は［小野89］で示した関係を一部拡張したものである）．

（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．

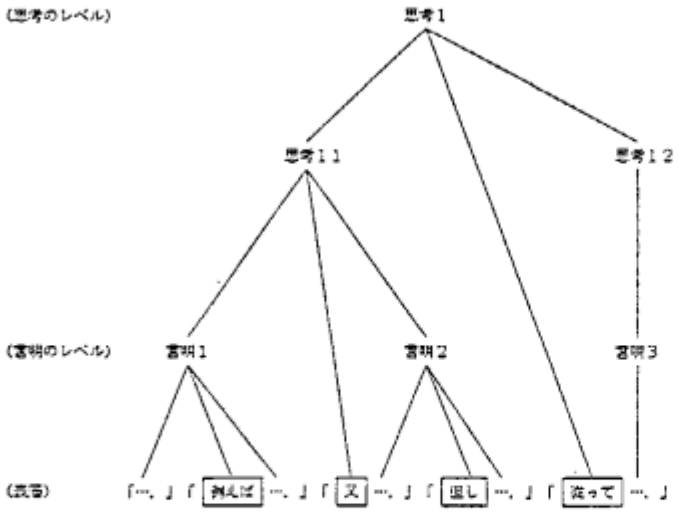

図3．1　言明，思考のレベルとその構造

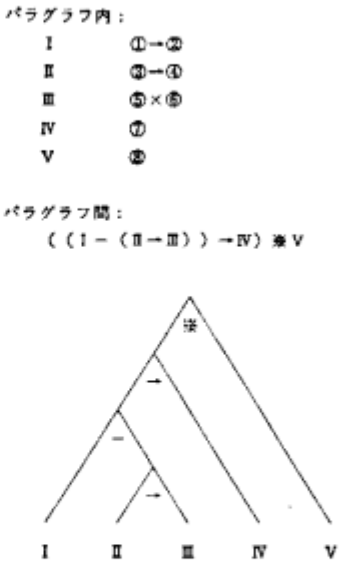

図3．2　文脈構造の例
（付録1の文章を構造化したもの．
文及びパラグラフの関係において，→は順接，
×は逆接，－は対比，※は指示である）

・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．

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"
⑤背景　　例：" 従来"
⑥定義　　例：" …とは"
⑦ハイライト　例：" …こそ"
⑧追加

（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．

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"
解答　⑩解答　例：" その答は…"

これらの関係は，基本的に２項関係であるが，言明レベルの「追加」や思考レベルの「並列」，「対比」，「指示」の関係に関しては，それらの関係にある文の連続は，２項関係の組み合わせとして考えるよりも，ｎ項関係として考えたほうが自然である．しかし，ここでは２分木を基本として考察を進める．

３．２　文脈構造の表現

２つの文又はその集まりの間の修辞的関係を決定していくと，結果として木構造が得られる．これが文脈構造となる．構造化の際は，まずパラグラフ内での文間の関係による構造化を行い，次にパラグラフ間で構造化を行うという形の２つのレベルで行っている．

図３．２は文脈構造表現の一例である．この例は，付録１に示すテキストの構造を抽出したものである．このテキストは，全８文，５パラグラフよりなっており，①～⑧は文の番号，Ⅰ～Ⅴはパラグラフの番号である．

４．構造抽出処理

上で述べたテキスト理解モデルに基づき，テキストの文脈構造を抽出する処理について述べる．

表層の表現から，文脈構造を抽出する過程は，次の３つのステップからなる．

１．単文化処理
２．文解析
３．構造化処理

４．１　単文化処理

単文化処理では，処理の単位となる文を認定するためのもので，接続詞もしくは接続助詞などで結合された文を，個々の文のレベルに分解する．構造化においては，分割される前の単位が構造として優先されるため，分割された個々の文が直接外の文と結び付けられることはない．しかし，論旨を展開する上で構造の類似性から，文間の関係の推定に利用することができる．

４．２　文解析

文解析においては，単文化処理で決定された文に対し，各文に現れる接続表現など，構造化を行う際に使用するデータを抽出する．現在は形態素解析の結果をもとに，以下に示すデータを抽出している．

前方の文との関係

主として文の先頭に現れる接続的表現を基に，先に述べた修辞的関係を決定する．当然のことながら，接続的表現がないために，関係が決定できない場合も多い．また，ここで取り出された関係は，構造化されるさい必ずしも直前の文との関係というわけではない．

主題

助詞の「は」に代表される話題提示表現をもとに，主題化されている名詞を抽出する．また，パラグラフ内の前方の文の中で主題化されている名詞を含むものを調べる．

文のタイプ

主節の述語などを参照して，次に示す７つのタイプに分類する．

理由文	あることがらに対する理由を述べた文．
規範文	あることがらに付随した義務などを述べた文．
問題提示文	あることがらを問題として提示する．
意見文	書き手のもつ認識や意見を述べた文．
背景文	近況など，ある時点での状況を述べた文
メタ文	図の参照を促したりする文
事実文	現在の状況や既に起こったことがらを述べた文．

述語と文末表現

主節の述語を取り出す．また，その付属語を調べ，テンスとアスペクトなどにより分類する．

4．3　構造化処理

前章で述べた思考制約や言語的手掛かりをもとに構造を絞り込むための知識を規則として表現しておき，それをもとに，文間の全ての可能な構造から不適当な構造を排除し，またより優先度の高い構造を取り出す．

4．3．1　思考制約による構造の絞り込み

（1）修辞的関係を用いた構造禁止規則

以下の構造を含む候補は，思考の流れとして考えにくいため，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　…

ここでXは，1つの言明あるいはそれらを修辞的関係で組み合わせたものを表す．これにより，次のような部分木を持つ文脈構造は棄却される（ここで"→"は直列型の関係の1つである順接を表す）．

（…　→　（文　→　…））

（2）修辞的関係を用いた構造優先規則

（2－1）以下の構造を含む候補は，論説文においてはあまり用いられないため，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　…
…，同示型関係　（　X　並列型関係　…

（2－2）以下のものはその構造の優先度を上げる．

左結合の構造の優先

論旨を展開する際は，その前で議論されていることがらに対して説明を付加したり，それに対する評価を行なうことで，論旨が展開されていく．したがって，その基本的構造は以下のような左結合の構造となることが多い（"？"は関係が不定であることを示す．以下の例でも同様）．そこで，特に修辞的な手掛かりのない場合には，デフォルトとして左結合の構造を優先する．但し，この構造の優先性は，修辞的関係による構造の制約に比べ明確な根拠のあるものではないので，次に示す言語的な手掛かりと共に使用する．

（（①　？　②）　？　③）

［小野89］で示したように，技術論文を対象として実験を行った結果，（1）のみでは，構造は1／3程度までしか構造が絞れないが，（2－1）まで用いると，1／5程度まで絞り込むことができる．しかし，これらの制約は，文間の修辞的関係を利用した制約であるため，接続的表現がない場合には，関係の抽出が困難であり，効果的な絞り込みができない．また，（2－1）による絞り込みでは，正解の構造の優先度を下げる場合があることが分かっている．

4．3．2　言語的手掛かりを用いた構造の絞り込み

構造の優先度を上げるものとして次のようなものがある．

＊主題化されている語句を含む文との接続関係の優先

主題を含む文と，その文の前方にあってその主題化されている語句を含む文との関係によって構成される構造を優先する．

＊文末表現または述語が同じ文を結ぶ構造の優先

文末表現が同じ文同士は，同じレベルの関係となる場合が多い．例えば，次のような文の並びの場合，（a），（b）2つの可能性があるが，文末表現の同一性から（a）の構造が優先される．

①～していた．
②これは，～である．
③～していた．

（a）　（（①　？　②）　？　③）
（b）　（①　？　（②　？　③））

文末表現そのものでないにしろ，テンスやアスペクトなどその一部が同じ場合にも文間の関係が同じレベルであることも多い．述語が同じ場合も同様である．

＊文のタイプによる構造の優先

論説文では，一般に，単にある事柄を述べた事実文や，背景を述べた文よりも，書き手の意見を述べた文のほうが重要度が高い．そのため，意見文がパラグラフの最後に現れている場合は，それ以外の文からなる部分構造を受けた形となっている構造を優先する．また，文のタイプが理由文である場合には修辞的関係が明示されていなくとも例示の関係として認識できるため，前の文との直接的な関係のある構造を優先する．

表５．１　分析結果
（カラム（１）～（３）の意味及び表中の各記号の意味は本文５．２節のこと）

パラグラフ番号	文数	（１）	（２）	（３）	
				(a)	(b)
1	3	2	1 ×		A ×
2	4	5	1 ○		A ○
3	5	14	2 ○	1 ○	
4	8	429	151 ○	30 ×	
5	4	5	3 ○	1 ○	
6	9	1430	317 ○	63 ×	
7	3	2	2 ○	1 ×	
8	3	2	1 ×		B ×
9	3	2	1 ○		C ○
10	6	42	37 ○	7 ○	
11	4	5	2 ×		A ×
12	3	2	1 ○		C ○
13	3	2	1 ○		A ○
14	4	5	3 ○	1 ○	
15	3	2	1 ○		C ○
16	3	2	2 ○	1 ○	
17	5	14	6 ×		C ×
18	7	132	9 ×		A ×
19	4	5	4 ○	1 ×	
20	4	5	6 ○	1 ×	
21	4	5	5 ○	1 ×	
22	4	5	5 ○	1 ○	
23	6	42	28 ○	6 ○	

ここにあげた規則は，思考制約による構造禁止規則に比べ制約力が弱いため，構造を優先させる規則であっても，それにより，あてはまる構造が直ちに決定できるわけではない．そこで，各優先規則に点数を割り当て，その規則を満足する構造に対し点数を加算するなどの方法をとる必要がある．

なお，ここでの構造化処理は，全ての可能な構造の中から明らかに不適当なもののみを削除し，優先度の高いものを選択することで，妥当性の高い候補から順に構造を得ることができる．但し，全ての可能な構造の数は，文の数が増加するに従って指数的に増加する（単純な２分木構造の場合，その数はカタラン数（［Church82］）として計算できる）ため，処理に用する時間的な制約から，このような単純な方法では，処理が難しいという問題がある．

５．分析例

５．１　実験方法

前章で述べた構造化処理により，テキストの構造化の実験を行った．解析の対象は，電子通信学会論文誌および東芝レビューに掲載された論文各１編におけるパラグラフの中で文の数が３文以上のものである．なお，文間の関係が明示されていないものに関しては，人間が関係を与え，全ての文間に関係を与えた状態で計算機により構造化を行った．

まず，文の並びに対して可能な構造を全て作成する．次に思考制約を用いて，４．３．１で述べた規則により，文脈構造として不適当な構造を削除する（（２－１）として優先度を下げるの規則に抵触する構造も，今回の実験ではこの段階で削除している），続いて言語的手掛かりに基づいて表現した規則を用いて，残った構造の間で優先度の高い構造を選択する．

５．２　実験結果

分析結果を表５．１に示す．まず，表のカラムにおいて「文数」は各パラグラフを構成する文の数である．また，カラム（１）～（３）は次に示す通りである．

（１）可能な全候補数

（２）思考制約を用いて抽出した候補数，及びその中に正解が含まれているかの判定
（○は正解が含まれていることを示し，×は含まれていないことを示す）

（３）言語的手掛かりの効果

（ａ）　記述が（２）と同じ形式のものは，言語的手掛かりを用いて絞った構造の数と，それらの中に正解が含まれているかの判定を示す．但しここでは，構造を絞る際，文の数が５文以下のものは，全候補の数が多くても１０個程度であり，構造を唯１つ選びだしそれを正解と比較している．また，文の数が６

文以上のものは，（2）の段階でも構造数が多く，この中から無理に1個選びだしても，それが正解となる可能性は小さい．そのため，現段階では（2）で得られた候補の1／5の個数を選びだし，その中に正解があるかを調べている．

（b） 記述がA～Cのものは，（2）の段階で既に構造が1個に決定されているか，2個以上であっても，それらの中に正解が無い場合に関するものである．Aは正解構造に対して言語的手掛かりの点でも正しく絞り込みを行っており，構造化にプラスの効果を与えていると判断できるものを示す．また，Bは逆効果となっているもの，Cは明快な判断が下せないものを示す．

表に示す実験結果から，言語的手掛かりは，構造の絞り込みに対して約半数のケースで有効であることがわかる（（2）から（3）で構造の数が減っているか，（3）で" A" と判断しているケースが過半数となっている）．そのうち，さらにその半数では，構造の絞り込みに対して実質的に機能している．また，残りの半数については，（a）思考制約での絞り込みによって構造が1個に決定されているため，実質的な効果はなかったものの，正解とした構造を裏付ける効果を与えている．（b）（2）の段階で不正解になっている場合でも，思考制約を緩めることで正解の構造を抽出できる可能性がある，ということが分る．このことから，文間の関係が明示されていない場合でも，それらを用いて構造の絞り込みがある程度可能であると考えられる．

5．3 考察

失敗例に関する分析結果より，言語的手掛かりの利用に関して以下の問題点が明らかとなった．

述語と文末表現を利用した構造の優先は，単一の文のレベルという極めて局所的な構造のみを調べ，そういった構造を部分構造として持つ候補の優先度を上げるという形で行っている．そのため，文の数が多くなると現れる，より大局的なレベルでの構造が認識できずに，正解となる構造を優先することができないという問題点がある．パラグラフ番号が4の結果は，この典型例である．これに関しては，単なる文の間だけでなく，その集まりからなる部分構造同士においても，その中心となる文間でこれらの関係を調べることで対処できる．

また，文のタイプに関しては，分析対象としたテキストでは，ほとんどが事実文の集まりであり，構造化に対する有効性は確認できなかった．しかし，対象を社説などのように，書き手の認識や意見を多く含んでいたり，対象となる読み手が不特定のため，背景や状況の記述を含むような場合には，構造化に利用できると考えられる．

さらに話題表現に関しては，現在，注目している文と，その文の前方にあって，その文で話題化されている語句を含む文との関係に着目して構造化に利用している．しかし，次のような場合，文②③が話題表現により共に①と関係があることを利用して，それらが並列の関係にあることまで検出する必要がある．

①～としてはAとBがある。
②Aは・・・。 ・・・。
③Bに関しては，・・・。

なお先に述べたように，文数が6文以上の場合は，可能な全候補数が指数的に増加するため，思考制約で絞り込みを行った時点での構造数も多く，今回の実験では言語的手掛かりを用いて，それらを1／5程度に絞り込むことを目標とした．最終的な構造の数をさらに減らすには，次のようなことが考えられる．

文の数が増加すると，パラグラフの中で述べている内容の違いから，なんらかのまとまりが感じられることがある．このようなまとまりは，現在構造化に利用していない同語反復や類語反復を手掛かりとして推定できると思われる．これにより，テキストをパラグラフの中とパラグラフ間の2つのレベルで構造化するのと同様に，パラグラフ内での構造化でも，部分的に構造化を行ってから全体の構造を認識することが可能となる．その結果，初めから一度にパラグラフ全体を構造化する場合に比べ，もとの候補の数を押さえることができ，最終的候補の数を減らすことができると思われる．

また，テキスト全体の構造化において，初めにパラグラフ内で構造化を行い，その後でパラグラフ間で構造化を行う方法では，次のようにうまく構造化出来ない場合がある（各例で，Ⅰ，Ⅱ・・はパラグラフの番号を，①，②・・は文番号である）．

（例1） Ⅰ ①～は3つある。②先ず第1は～である。
Ⅱ ③第2は～である。④これは・・・。
Ⅲ ⑤第3は～である。

この例では，内容から考えると，文②と文③④，文④が並列の関係にあり，それらを文①がまとめる形となる．しかし，初めにパラグラフ内で構造化してしまうと，このような構造は得られない．

（例2） Ⅰ ①～はAとBから構成される。

②このＢには，Ｘに関する情報を記述しておく。
Ⅱ　③例えば，～の場合，Ｂの内容は次のようになる。
④・・・。

この例では，パラグラフⅡで述べられていることがらは，文②の例であって，パラグラフＩで述べられていることがら全体の例ではない．しかし，パラグラフ内で先に構造化すると，後者の解釈になってしまう．

6．おわりに

本稿では，テキスト理解と文脈構造との関係について整理し，テキストの生成と理解の観点から文脈構造の位置付けを明らかにした．次に計算機によりテキストの文脈構造を得るためのテキスト理解のモデルを示した．このモデルでは，接続表現により判定される文間の修辞的関係を用いた思考上の制約や，話題表現や文末表現，文のタイプなどの言語的手掛かりを用いて，構造の適切性や優先性を調べ，テキストの構造化を行う．続いて，このモデルに基づき上記の手掛かりを具体的に検討して，構造抽出の方法を示した．さらに，構造化の実験を行って，これら言語的手掛かりの利用により，思考制約をのみ用いた場合に比べ，さらに構造を絞り込むことが可能であることを確認した．しかしながら，構造を一つに決定するためには，現在の処理でも不十分であり，そのためには，非言語的知識の利用が不可欠である．

構造化の手掛かりの１つに類語の反復がある．同じ語又はその一部が繰り返して用いられる同語反復では，語の概念まで考慮する必要はないが，類語の反復では概念の階層構造を参照する必要があり，非言語的知識を利用する最も簡単な手掛かりといえる．また，文の同義性を認識することも構造化において重要な手掛かりであるが，これを認識するには，構造化された概念の類似性を判定できなければならない．今後，これらの観点から始めて，知識処理まで含めた構造抽出モデルを検討していく．なお本研究は，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．

参考文献

[Church82] Church, K. and Patil, R. :'Coping with syntactic ambiguity or how to put the block in the box on the table, ACL 8(3-4), pp. 139-149, 1982.

[Cohen87] Cohen, R. : 'ANALYZING THE STRUCTURE OF ARGUMENTATIVE DISCOURCE', Computaional Linguistics, Vol. 13, No. 1-2, 1987.

[Grice75] Grice, H. P. : 'Logic and Conversation', Syntax and Semantics, Vol. 3, Speech Act, Seminar Press, pp. 41-58, 1975.

[Johnson-Laird83] Johnson-Laird, P. N. :'Mental Models', Hardvard University Press, 1983.

[木下88] 木下，佐野，浮田，住田，天野：" 文脈理解のための知識の表現と推論"，Proc. of LPC'88, pp. 205-215, 1988.

[Leech83] Leech, G. N. :'PRINCIPLES OF PRAGMATICS', Longman group limited, 1983.

[Mann87] Mann, W. and Thompson, S: 'Rhetorical Structure Theory: A Framework for the analysis of Texts', USC/Information Science Institute Research Report, RR-7-190, 1987.

[永野86] 永野賢：" 文章論総説－文法論的考察－"，朝倉書店，1986.

[小野89] 小野，浮田，天野：" 文脈構造の分析"，情報処理学会　自然言語処理研究会 70-2, 1989.

[Rumelhart75] Rumelhart, D. E. : 'Notes on a schema for stories', in D. G. Bobrow and A. M. Collins (Eds.), Representation and understanding, Academic Press, 1975.

[Scha88] Scha, R. and Polanyi, L. ; 'An Augmented Context Free Grammar for Discource', Proc. COLING-88, pp. 573-577, 1988.

[Schank77] Schank, R. C. and Abelson, R. P. ; 'Scripts, plans, goals, and understanding', Lawrene Erlbaum Associates, 1977.

[所86] 所一哉：" 日本語　思考のレトリック"，匠出版，1986.

[辻井88] 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.

[Ukita88] Ukita, T., Sumita, K., Kinoshita, S., Sano, H. and Amano, S. :'PREFERENCE JUDGEMENT IN COMPREHENDING CONVERSATIONAL SENTENCES USING MULTI-PARADIGM KNOWLEDGE', Proc. of FGCS'88, pp. 1133-1140, 1988.

[浮田89] 浮田，小野，住田，木村，木下，天野：”談話文脈構造に関する考察”，ディスコースと形式意味論ワークショップ，pp. 91-100，1989.

付録1　文章例

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）