ICOT Technical Memorandom: TM-1162

TM-1162

文の連接関係解析に基づく文章構造解析

福本　淳一、安原　宏（沖）

March, 1992

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03)3456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

文の連接関係解析に基づく文章構造解析

福本 淳一　　安原 宏

沖電気工業(株) 総合システム研究所

概要

本稿では、新聞社説記事の分析から得られた結果を基にした文章構造モデル、及びそのモデルに基づく言語仕様について述べる。また、この言語仕様に基づいて記述された規則を用いた文章構造解析システムの構造解析手法についても述べる。本システムでは、まず、文章中の連接２文間の関係の解析を行い、その関係を基に文のグループ化を行う。そして、生成されたグループ間の関係を解析し、文章構造フレームへの埋めこみを行うことで文章全体の構造を得る。これまでに約１７０の規則を記述し、これを用いて新聞社説記事の構造解析を行い、動作を確認している。

Text Structure Analysis based on Sentence Cohesion

Jun-ichi FUKUMOTO, Hiroshi YASUHARA

Systems Laboratory, Oki Electric Industry Co., Ltd.
11-22, Shibaura 4-chome, Minato-ku, Tokyo 108, Japan

Abstract

In this paper we propose Japanese text structure analysis method based on the analysis of the relationships between sentences. At first, each relationship between a sentence and its neighbor in a text will be analyzed. Secondly some groups of sentences will be generated with using these sentence relationships. Then the relationship between these groups will be analyzed and fill the groups in a text structure frame. We present the text structure model and rule formalism to analyze the Japanese text structure. We show an example of the text structure using about 170 rules on the analysis of some Japanese texts.

1 はじめに

論説文は、ある事柄についての書き手の考えや意見など、書き手の主張を述べることを目的とした文章であり、このような文章においては、書き手の主張を読み手である読者に伝えるために論旨が展開されている。この論旨の展開構造が、文章の構造であり、書き手が文章において主張したいことであると考えられる。これまで我々は、論説文として新聞の社説記事をとりあげ書き手の主張という観点から文章の構造化についての分析を行ってきた[1][2][3][4][5]。このような構造化のため、文章中の各文を書き手の主張に基づいてタイプ分類し、各文のタイプ情報、主題提示情報、文章中に表れる接続詞等の機能語の情報を用いることで文章を木構造の形式に構造化してきた。しかし、このような構造では、文章中に表れる現象が十分に表せておらず、文章構造を解析するための規則も形式化されていなかった。

我々は、文章の構造化のためのモデル、及びそのモデルに基づく言語仕様を定義し、これを用いて記述した規則により文章の構造解析を行うシステムを開発した。文章構造の解析は、文章中の連接2文間の関係の解析を行い、その関係をもとに文のグループ化を行う。そして、各生成されたグループ間の関係を解析し、文章構造フレームへ埋め込むことにより文章全体の構造を得る。

以下、2章で文章構造のモデルについて述べ、3章で文章構造の解析のための記述規則の言語仕様について述べる。そして、4章で文章構造解析システムの概要、及び解析例を示す。

2 文章構造のモデル化

これまでに我々は、文の連接関係を一方が主で他方が従であるような主従関係として捉えることで文章の構造化を行ってきた。しかし、実際の文章中には、主従を決定しにくいものがいくつか存在する。文章中で、ある事柄について述べられた文が順に列挙される場合、これらの文間には主従の関係にあたるものは考えられない。以下の例で、数字はパラグラフ番号 - 文番号を示す。

1-1 ジョギング中の突然死や社長の急死などの不幸な事件が、このところ目立っている。
1-2 産業構造が変わり技術革新が進んで、働く人のストレスもつのってきた。
1-3 高齢化への歩みが速まるなかで、働き盛りの中高年の健康管理が特に重要な問題になっている。

（昭和62年10月2日付け朝日新聞社説記事より）

また、「第1に」「第2に」「第3に」のような表現を用いることで、いくつかの文のまとまりが序列的に並んでいるものもある。

5-1 第1に、水の節約が、水道料金の値上げにつながらぬようにしなければならない。
8-1 第2に、給水配管の漏水対策を積極的に進めてほしい。
9-1 第3に、最も重要なのは、今回の水不足が、東京集中によってひき起こされた構造的な問題だ、という認識を持つことである。

（昭和62年8月30日付け朝日新聞社説記事より）

このように連接関係の中で主従関係がなく、文や文のまとまりが連続して述べられているものを1つのまとまりとして表現することが考えられる。このまとめる操作をグループ化と呼ぶことにする。

また、文の連接関係を主従関係と捉えることで得られた文章構造に対しても、このグループ化を用いることが考えられる。これは、文章中である内容について述べられているいくつかの文を、1つのまとまりとして表現することである。

以上のように文章構造を表現するためには、各ノード間の連接関係を主従関係と認識するだけでなく、いくつかの文をグループとしてまとめて表現することも必要である。このグループ化により、同じ話題について述べられている文や連続してある事柄が述べられている文を1つのまとまりとして表現できる。

2.1 文章構造モデル

文章構造をモデル化するため「。」で区切られた1文を1つのノードとして表し、これらの文間に主従関係が認められる時、その関係をノード間のアークで表す。このとき、アークは、従となるノードから主となるノードへの向きがある。また、いくつかのノードをまとめてグループ化されたものも1つのノードとして扱う。本文章構造モデルで扱うノードは、文を表すSノード (sentence node) とグループを表すGノード (group node) の2種類がある。ノードの性質はノードの属性として表現され、ノードの種類も属性として表される。

(1) Sノード
文章中の各文をノードとして表したものである。
(2) Gノード
幾つかのノードをグループとしてまとめて1つのノードとしたものであり、次の2つのタイプがある。このタイプはノードの属性として表される。

○連続タイプ
主従の関係がなく連続して述べられているものや序列的に述べられているものをグループとして表現したノード
○スコープタイプ
連接関係が存在するノードのうち同じ話題に関するものをまとめてグループとして表現したノード

アークは、ノード間に主従の関係が存在する場合、これらのノードを結んだものである。このときノード間の関係で主になる側のノードを主ノード、もう一方のノードを従ノードと呼び、従ノードから主ノードへの向きがある。各ノードからは1つのアークが出ており、複数のアークが入ってくる。アークは、ノードと同様に属性をもつ。

Gノードの各タイプを表現したものを図1に示す。図中、ノードは「○」で、アークは「→」で表されている。

図1　Gノードの種類

文章構造は、ノード(Sノード, Gノード)間にアークを張ったり、ノードをグループ化することで構成されていく。アークを張るための条件は、ノードに付与された属性情報や他のアーク情報、Gノードの場合にはその要素のとなっているノードの属性情報によって決まる。このモデルを用いて表現された文章構造例を図2に示す。

3 文章構造解析規則

ここでは、2章で述べた文章構造のモデルに基づき、文章構造を解析するための文章のデータ構造、及び構造化規則の言語仕様について述べる。

3.1 データ構造

文章構造を解析するために用いるデータ構造としては、文章構造を表現するための文章構造データと構造化のために必要となる情報を記述するためのファクト型データの2種類がある。

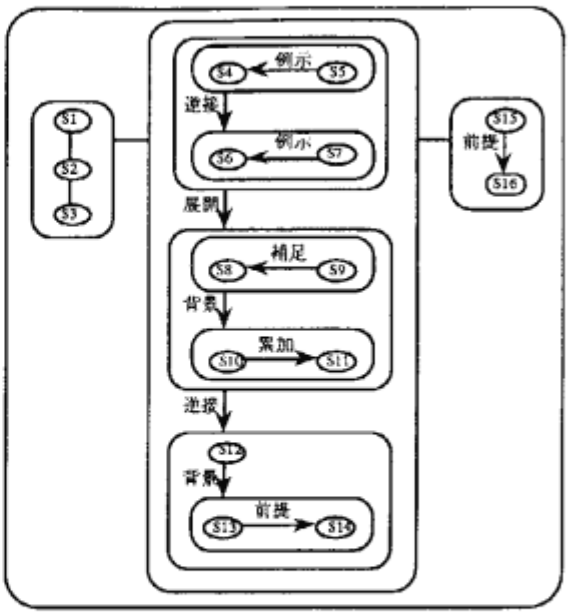

図2　文章構造の概略

(1) 文章構造データ

文章構造データには、ノード及びノード間に存在するアークの情報がある。ノードの情報としては、ノード名、ノード属性リスト、親ノード名、要素ノードリストがあり、1つのリストとして表現されている。親ノード名は、ノードがGノードの要素である場合、そのGノード名である。要素ノードリストは、ノードがGノードの要素である場合、その要素ノードのリストである。ノード属性リストは、属性名、値のペアのリストである。ノードs1,s2をグループ化したものがg3である例を以下に示す。

[例]

```
[g3,[[node_type,'g_node'],[grp_type,'スコープ']
    ,[grp_top,'s6'],[stype0,'主張文'],[stype1,'義務文']]
    ,all_sentences,[s1,s2]]
[s1,[[node_type,'s_node'],[stype0,'主張文']
    ,[stype1,'判断文']],g3,[]]
[s2,[[node_type,'s_node'],[stype0,'叙述文']
    ,[stype1,'状態'],[topic,'yes']],g3,[]]
```

アークの情報としては、アーク名、アーク属性リスト、始点ノード名、終点ノード名があり、1つのリストとして表現されている。アーク属性リストも、属性名、値のペアのリストである。以下にノードs1, s2の間に存在するアークarc5の例を示す。

[例]

```
[arc5,[[name,'前提'],[direction,'xy']],s1,s2]
```

(2) ファクト型データ

文章の構造に関する情報の他に、任意の情報をフ

ァクトの形で記述することができる。これにより、文章構造を作るために必要な文章中の任意の情報を記述することができる。ファクト情報としては、ファクト名とファクトのリストとして表現される。ファクトは、ファクトを表す述語名とその述語のもつ引数の値によって表される。以下にその例を示す。

[例]

```
[[f1,topic_word(s1,['調査','では'],'では','主文')]
[f2,topic_word(s1,['解消策','は'],'は','主文')]
```

3.2 言語仕様

本言語で記述された1つのルールは、ルール名、ルールグループ名、パターン部、条件部、実行部の各部から構成されている。各ルールは1つのルール名をもつ。ルールはいくつかをまとめてグループ化されており、ルール名の後にそのルールグループ名を書くことで表す。

```
<ルール名> : <ルールグループ名>{
        <パターン部> <条件部> <実行部> }
```

このルールのグループ化により、ルールの発火をそのグループ内のルールだけに限定することが可能となる。

パターン部で記述されたノード、及びアークに関する条件を満たすものが文章構造データに存在し、条件部に記述された全ての条件が真の時、ルールは発火される。そして、実行部により文章構造データ、及びファクト型データが書き換えられ、再びルールとパターンマッチにいく。そして、適用できるルールが存在しなくなった時点で実行は終了する。以下では、ルールの各部について述べる。

(1) パターン部

パターン部では、文章構造データとのパターンマッチにより、処理の対象になるノード、及びアークの指定を行う。また、ノード、アーク属性の条件チェックも行う。文章構造データとのパターンマッチは、次のように記述する。

```
pattern: <ノード名> : <ノードパターン>;
```

このとき、<ノード名> でGノードを指定し、<ノードパターン> でそのGノードの要素ノードの指定を行う。また<ノードパターン> 中のGノードについても同様に

```
<ノード名> : <ノードパターン>;
```

と記述することにより、パターンマッチが可能である。<ノードパターン> 中の両端には「*」ノードを記述することにより、0個以上の任意の数のノードとマッチすることができる。例えば、パターン文が、

```
top:[* x1 x2 x3 *];
```

のとき、文章構造データして [a b c [d e f g] h] が存在した場合、各ノード [x1, x2, x3] は、それぞれ [a b c], [b c [d e f g]], [c [d e f g] h], [d e f], [e f g] とマッチする可能性がある。

文章構造データの任意の2つのノード間にアークが存在すかどうかの照合は、述語 arc を用いて記述する。述語 arc は2つのノード名引数としてを持ち、それらの間にアークが存在することを示す。但し、それらのノードのうち少なくとも一方は pattern記述で表れたノードでなければならない。例えば、ノード x1 から x2 に向かうアークが存在することを arc(x1,x2) と表す。また、存在しないことを先頭に「!」を付け !arc(x1,x2) と表す。

文章構造データ内で定義されたノード、又はアークの属性に関する記述は「ノード名#属性名」「アーク名#属性名」の形式で記述し、その値の照合を式で記述する。オペレータがなく「ノード名#属性名」の形式のみの場合は、その属性が値をもつことを示す。

```
[例]    x1#stype == '間掛文';
        x1#stype == x2#stype;
        x1#stype;  (ノードx1のstype属性が値をもつ)
        arc(x1,x2)#name == arc(x2,x3)#name;
```

(2) 条件部

条件部では、ファクト型データを記述することにより、そのデータが存在するかどうかの条件の記述を行う。データが存在しない場合は、先頭に「!」を付与することで表す。

```
[例]    f1;topic_word(s1,_,'は',_);  (f1はファクト番号)
        !topic_word(s1,_,'は',_);
```

また、任意の関数（外部関数）を記述することも可能である。この関数は、true 又は false の結果のみを返す関数として記述される。関数名には、先頭に「@」を付けて表す。

```
[例]    @:function(#function,x1,x2);
```

(3) 実行部

実行部では、文章構造データ、及びファクト型データの書き換えに関する記述を行う。文章構造データに関する処理には、以下のものがある。

○ノードのグループ化

```
make_group(<ノードリスト>,<Gノード名>);
```

○グループノードの削除

delete_group(<Gノード>);

○ノードのグループへの追加

add_node(<ノード名>,<Gノード名>);

○ノードのグループからの削除

remove_node(<ノード名>,<Gノード名>);

○ノード及びアーク属性に関する処理

set(<ノード属性記述>,<値>);

set(<ノード属性記述>,<ノード属性記述>);

○アークの追加

make_arc(<ノード名>,<ノード名>);

○アークの削除

delete_arc(<ノード名>,<ノード名>);

ファクト型データの削除/修正は、条件部で記述されたファクト番号を指定することにより行う。ファクトデータの追加は、追加するファクトを記述することにより行う。

○ファクトデータの追加

add(<ファクト型データ>);

○ファクトデータの削除

delete(<ファクト番号>);

○ファクトデータの修正

modify(<ファクト番号>,
<ファクト型データ>);

各ルールは、ルールグループ名を持っており、これによってルールのモジュール化を可能にしている。

change(<ルールグループ名>);

実行部中には if 文によって条件分岐を記述することも可能である。条件節中の条件文には、&& (and)、|| (or) によって式を結合することができる。ここで論理式の評価の順序のあいまい性をなくすため、1つのオペレータで結ばれたものは ( ) でまとめるものとする。条件文の後には、1つの実行文のみを記述することができる。2つ以上の実行文を記述する場合は、それらを { } でまとめて記述する。

(4) 記述例

本言語仕様に基づいて記述した文章解析規則の記述例を以下に示す。この規則のパターン部では、連続する2つのノードがそれぞれ主張文、叙述文であり、叙述文のみに主題提示があり、それらの間にアークがないことを表す。条件部では、叙述文の主題提示語が主張文中に表れないことを示す。実行部では、それらの間に転換関係のアークを張ることを示す。

```
rule1:sent_sent {
    pattern: top:[ * x y * ];
      x#stype0 == '主張文';
      y#stype0 == '叙述文';
      x#topic != 'yes';
      y#topic == 'yes';
      !arc(x,y);
    condition:
      topic_word(y#name,WORD,_,_);
      !rep_word_rel(x#name,y#name,WORD);
    action:
      make_arc(x,y);
      set(arc(x,y)#name,'転換');
}
```

4 文章構造解析システム

現在、以上のモデルに基づいて記述された規則を用いることで、論説文等の文章の構造解析を行うシステムの開発を行っている。図3にシステム構成を示す。

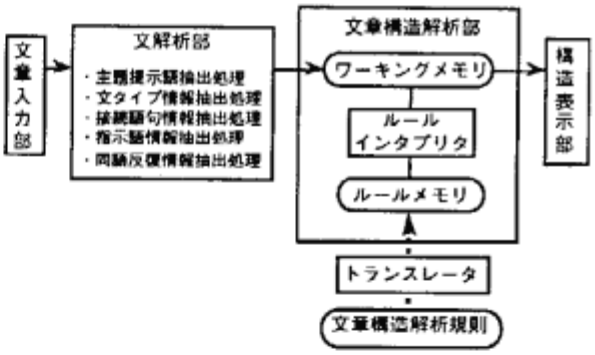

図3　文章構造解析システム構成図

4.1 文解析部

文解析部では、文章中の各文についての形態素解析結果をもとに、文章の構造解析のための入力となる情報として、主題提示語、文タイプ、接続語句、指示語、及び同語反復の情報の抽出を行う。これらの抽出された情報は、文章構造データ、及びファクト型データとして出力される。

(1) 主題提示語抽出処理

主題提示語抽出処理においては、文中で主題提示語となる語句の抽出を行う。一般に主題提示語は自立語に助詞「は」が付属することで示される。しかし、「特には」「～とはいえない」のように主題提

示の機能を持たないものも存在する。そこで、新聞社説記事を用いた調査から、以下のものを主題提示語として抽出した。

名詞+は　　名詞相当語+は
名詞+においては　　名詞+については
名詞+としては　　名詞+にとっては
名詞+では　　名詞+とは
名詞/名詞相当語+には

また助詞「も」が付属する語句についても「は」の場合と同様の基準で抽出した。

(2) 文タイプ情報抽出処理

文章を構成する各文は、書き手の考えや意見などの書き手の主張が表れている文（主張文）とそのような意見を表すために必要な文（叙述文）とに分類できる[1]。文タイプ情報抽出処理においては文章中の各文の文末表現を解析することにより、主張文については、[問掛文, 断定文, 推量文, 要望文, 判断文, 意見文, 理由文, 義務文]の8つに、叙述文については、[可能, 伝聞, 様態, 叙述, 存在, 継続, 状態, 使役]の8つに分類した[2]。また、文のテンス情報 (現在, 過去) についても抽出した。

(3) 接続語句情報抽出処理

接続語句情報抽出処理においては、文中で接続詞等の文間の関係を決定する語句、及び「第1に」のように文章の流れを決定する機能を持つ語句の抽出を行う。接続詞情報としては、品詞が接続詞であるものを抽出し、接続的な機能を持つ語句としては、新聞社説記事の分析で得られた結果をもとに抽出するものを決定した。

(4) 指示語情報抽出処理

指示語情報抽出処理においては、文中の指示代名詞、及び指示連体詞を含む文節を抽出する。但し、人称代名詞の1人称、2人称のものや「その半面」「このところ」等は抽出しない。後者については、新聞社説記事の分析で得られた結果をもとに抽出しないものを決定した。

(5) 同語反復情報

同語反復情報抽出処理においては、文章中で繰り返し用いられている名詞語句を抽出する。また、反復語に対して修飾語が掛かっているかどうかの情報も抽出する[3]。

4.2 文章構造解析部

文章構造解析部では、文解析部の出力である文章中の各文に関する主題情報、文タイプ情報等の情報を入力とし、記述された文脈規則にしたがって文章の構造解析を行う。文解析部の出力データは、ワーキングメモリにセットされ、ルールメモリ中の規則を適用しながらワーキングメモリの内容を順次書き換えていく。適用する規則がなくなった時点で、ワーキングメモリの内容が構造表示部によって出力される。

文章構造の解析は、まず、文章中の隣接文間の関係の解析を行い、文同士の連接関係の解析を行う (phase1)。次に、解析された文の連接関係をもとノードのグループ化を行い、文章全体をいくつかのグループノードに構造化する (phase2)。最後に、文章構造フレームに従って、文章を構成するために必要な要素をグループノードより認識し、文章の構成要素以外のノードのまとめ上げを行い、文章全体を構造化する (phase3)。

(1) 隣接文間解析処理 [phase1]

隣接文間の関係は、主として新聞社説記事の解析結果から得られた文のタイプ間の連接関係によって決定する。但し、連接文間に接続詞情報が存在した場合、また、指示語情報や主題提示情報などを用いて記述された規則に適用した場合、そちらを優先して連接関係を決定する。連接関係としては、次の17の関係を設定した。また、これらは2文間の主従関係によって以下の3つに分類されている。

○前文が主 (アークの向きは後文から前文)
例示…後文が前文に対する例示的内容となっている
補足…後文が前文に対する補足的内容となっている
呼応…後文が前文の問掛文と呼応関係になっている
○後文が主 (アークの向きは前文から後文)
背景…前文が後文に対する背景的内容となっている
前提…前文が後文に対する前提となっている
根拠…後文が前文に対する根拠となっている
結果…後文が前文に対する結果的内容となっている
○主従の関係のないもの (アークの向きはない)
逆接…前文と後文で反する内容が述べられている
転換…前文と後文で述べられている内容が変わる
序列…後文で序列的接続表現が用いられている
累加…前文に対して後文で付け加わる内容が述べられている
反復…前文に対して後文で内容の反復が行われている
並列…前文と後文で並列的内容が述べられている
対比…前文と後文で対比的内容が述べられている
継続…連続する叙述文において述べられている話題が

継続している
展開…指示語等が用いられることで述べられている内容が展開している
連係…前文と後文で関連のある内容が述べられている

(2) ノードグループ化処理 [phase2]

ノードグループ化処理では、phase1で解析された文の連接関係の中で結び付きの強いものをグループとしてまとめる。残った連接関係については、ノードの種類、及び連接関係名によってトップノード以下の全てのノードがグループノードになるまでグループ化を行う。また、生成されたいくつかのグループノード間の関係についても、そのグループの主ノードの情報等を用いることでグループ間関係の解析処理を行い、さらにグループノードのまとめる。

(3) 文章構造認識処理 [phase3]

文章構造認識処理においては、文章を構成するために必要な要素を記述した文章構造フレームを用いて、その要素となるグループノードを認識する。そして、文章の構成要素以外のノードをまとめたうえで、文章全体の構造を認識する。現在認識できる文章構造としては、文章を [序論、本論、結論] ととらえる規則のみであり、今後、phase2までの解析処理結果を調査することで、規則の拡張を行う必要がある。

4.3 文章構造解析例

本言語仕様に基づき文章の構造化のための規則の記述し、PSI-II 上で開発した言語処理系を用いて新聞社説記事の構造化を行った。新聞社説記事について構造解析を行った例を図4に示す。なお、図中の数字はパラグラフ番号_文番号を示し、[主] はその文が主張文であることを示す。また、[g] はグループノードを示す。

5 おわりに

本稿では、文章を構造化するため、新聞社説記事を用いた分析から得られた結果を基にした文章構造化のためのモデル、及びそのモデルに基づく言語仕様について述べた。そして、これを用いて記述した規則により文章の構造解析を行うシステムの解析手法について述べた。

今後の課題としては、文章全体の構造を認識するための文章構造フレームを拡張し、この構造化規則をいくつかの文章に適用することで、規則の評価、改良を行う予定である。また、構造化のための言語仕様の改良もある。

【謝辞】

本研究は第5世代コンピュータプロジェクトの一環として ICOT からの委託で行われたものである。研究を進めるにあたり、有益な助言を下さいましたICOT第6研究室田中室長、及びICOT NLUワーキンググループの皆様に感謝します。また、研究協力頂いた(株)沖テクノシステムズラボラトリの柴田、田中、斎藤各氏に感謝します。

【参考文献】

[1] 福本 淳一 : "筆者の主張に基づく日本語文章の構造化", 情報処理学会自然言語処理研究会 78-15, 1990.

[2] 福本, 安原 : "日本語文章の構造化解析", 情報処理学会自然言語処理研究会 85-11, 1991.

[3] 柴田, 田中, 福本 : "新聞社説記事における照応現象", 情報処理学会第40回全国大会, 1990.

[3] 田中, 柴田, 福本 : "文章構造解析システムにおける同語反復解析処理", 情報処理学会第43回全国大会, 1991.

[5] 斎藤, 柴田, 福本 : "文章構造化のための文の連接関係の解析", 情報処理学会第43回全国大会, 1991.

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

[7] 市川 孝 : "国語教育のための文章論概説",教育出版, 1978.

[8] 木下, 小野, 浮田, 天野 : "日本語テキスト理解における文脈構造抽出法",「談話理解モデルとその応用」シンポジウム, pp.125-136, 1989.

[9] Grosz,B., Sidner,C.L. : "The structures of Discourse Structure", Technical Report, CSLI, CSLI-85-39, 1985.

[10] Mann,W.C., Thompson,S.A., : "Rhetorical Structure Theory: Description and Construction of Text Structure", Proceedings of the Third International Workshop on Text Generation, 1986.

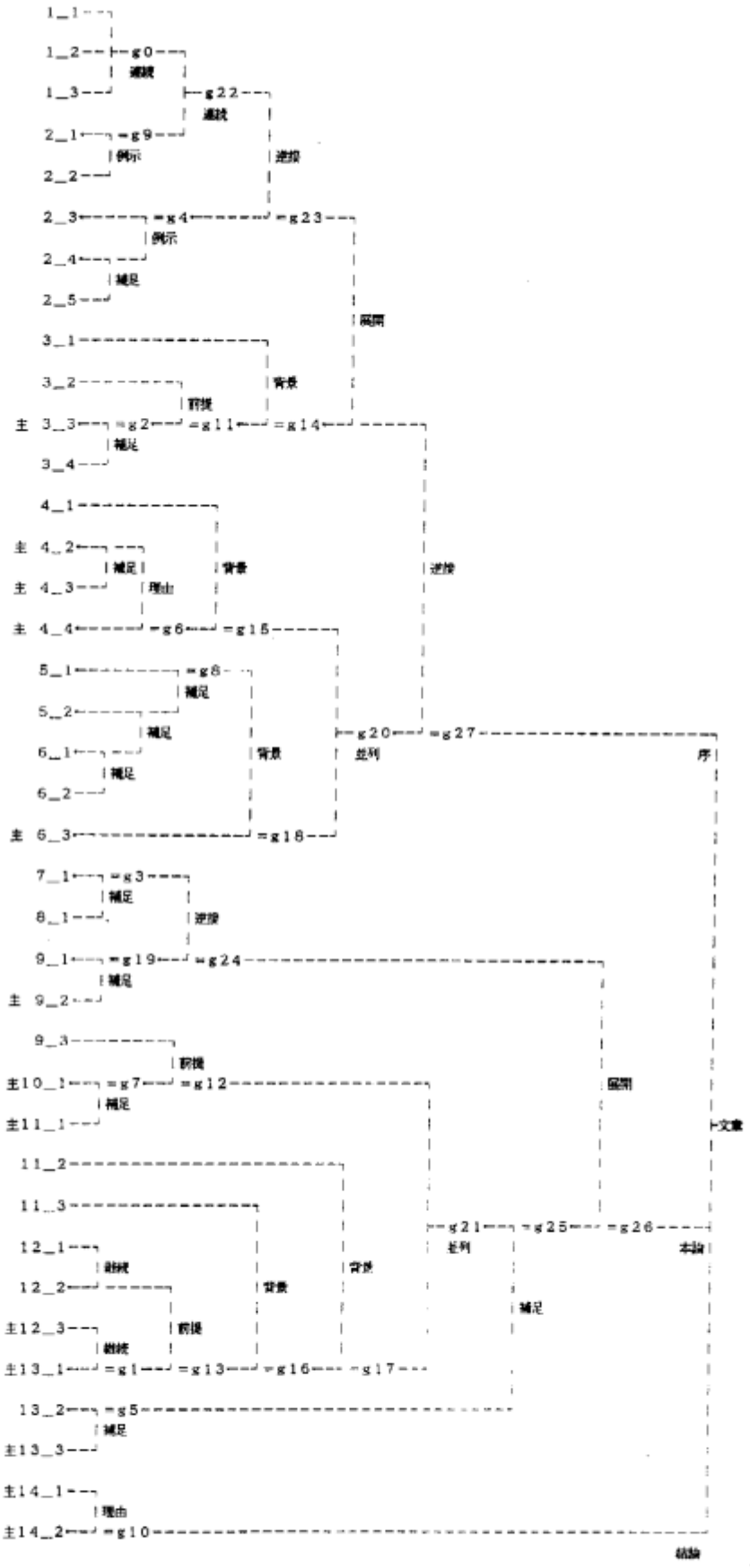
1_1
1_2
1_3
2_1
2_2
2_3
2_4
2_5
3_1
3_2
主 3_3
3_4
4_1
主 4_2
主 4_3
主 4_4
5_1
5_2
6_1
6_2
主 6_3
7_1
8_1
9_1
主 9_2
9_3
主10_1
主11_1
11_2
11_3
12_1
12_2
主12_3
主13_1
13_2
主13_3
主14_1
主14_2
g0
g22
g9
g4
g23
g2
g11
g14
g6
g15
g8
g20
g27
g18
g3
g19
g24
g7
g12
g21
g25
g26
g1
g13
g16
g17
g5
g10
連続
例示
逆接
補足
同開
前提
背景
理由
並列
序
主文
結論

図4　文章構造解析例