自然言語処理　36－1
（1983. 3. 14）

# 日本語文係り受けの分析・解析・学習

白井克彦　　林　良彦　　平田裕一

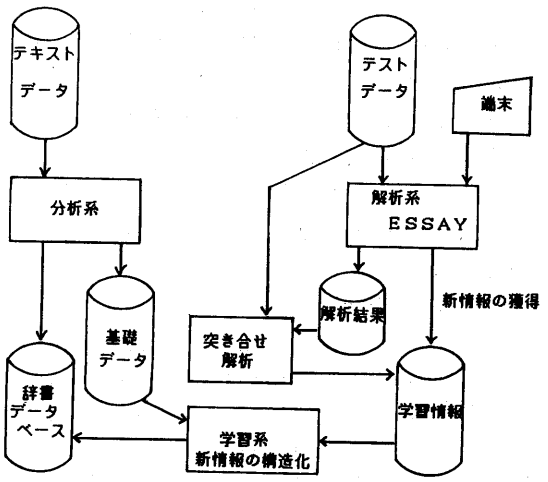
（早稲田大学　　理工学部）

## 1．はじめに

複雑でかつ多様な自然言語の現象に対処可能な処理システムを実現しようとする際，それに必要な多様なレベルの知識をどのようにして獲得し，利用するかについて，多くの有効な処理モデルが提案されているが，なお大きな問題である．〔1〕

我々は，人手により加工された形で蓄積された実テキストデータに対する分析により直接的に初期的な知識構造を構成し，これを文解析に使用することにより，この知識構造の構成手法を評価するとともに，この過程を通して得られる情報を，知識構造に反映させるというアプローチを日本語文の係り受け構造に適応してきた．〔2〕～〔4〕　本アプローチの概要をＦｉｇ．1に示す．

ここでいう知識とは，対象世界における言語要素の使われ方に関するものであり，辞書データベースとして記述的な形で構造化される．このように，使われ方に注目して，知識の獲得・構造化を行なうため，対象世界における拘束を緩やかに含み，かつインクリメンタルに成長しうる知識構造の構成が可能となる．

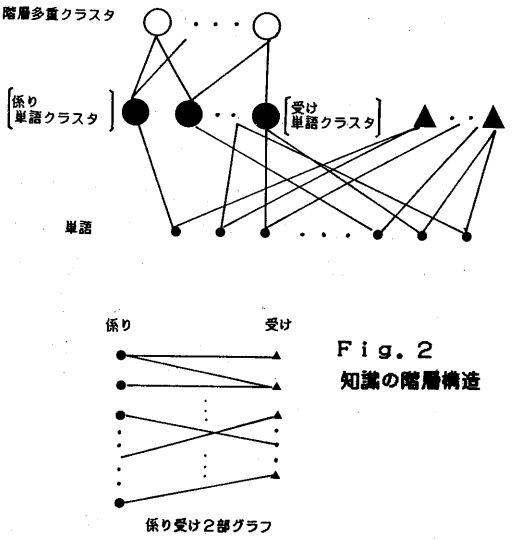



Ｆｉｇ．1　アプローチの概要

## 2．初期辞書データベースの構成

初期辞書データベースは，実テキストデータの分析により構成されるものであり，以後の成長に対してその核となるものである．

本辞書中では，分析対象の文中に存在する係り受け関係を構成する単語間の依存関係は，単語の集合であるクラスタ間の係り受け可能関係として構造化される．この知識構造をＦｉｇ．2に示す．以下，実データの分析により，この知識構造を構成する手法について述べるが，その概要をＦｉｇ．3に示す．




Ｆｉｇ．2
知識の階層構造

### 2．1　対象テキスト

含まれる単語数が比較的少ない，複雑な文が少ない等の理由により，小学校の教科書を対象とし，係り受け構造等を付加し約２５００文を登録した．

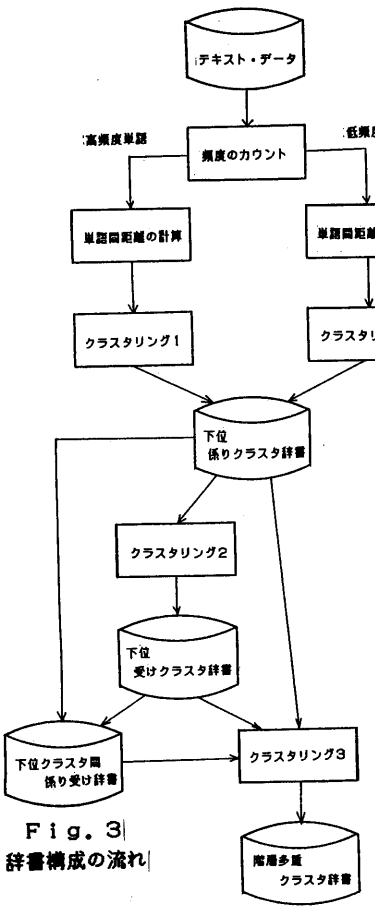



Fig. 3
辞書構成の流れ

## 2．2　単語クラスタ辞書の構成

テキストデータ中に存在する係り受け関係を，Fig. 2に示したように構造化することにより，テキストデータ中には実際には存在しない係り受け関係をも処理しうるという抽象化の効果が期待できる．単語は，1つの係りクラスタ，受けクラスタ及び複数の階層多重クラスタに属することになる．

### 2．2．1　単語間距離行列の計算

クラスタリングにより単語分類を行なうためには，単語間の距離を計算しておくことが必要となる．そこで係り単語に対しては，同じ付属語を介して同じ単語により多くの頻度で係っている単語間の距離が近いと考える事により，単語A，B間の距離を次の様に定義し，計算を行なう．

$$D(A,B) = 1 - \frac{\Sigma\,[\,M(A,f,a) + M(B,f,a)\,]}{\Sigma\,F(A,c) + \Sigma\,F(B,d)} \qquad (1)$$

ここで

$M(\alpha,\beta,\gamma)$：単語$\alpha$が付属語$\beta$を介して単語$\gamma$に係かる頻度

$f \in K(A) \cap K(B)$

$K(a)$：単語aにつく付属語の集合

$F(a,b)$：単語aから単語bへの係り受け頻度

である．

### 2．2．2　単語の出現頻度の考慮

現在テキスト中に出現する係り単語数は約1500，受け単語数は約1000であるが，テキストに1度しか現れない単語は，係り単語で約40%，受け単語で約44%にものぼっている．このような出現頻度の低い単語，つまり情報量の少ない単語は，クラスタリングの際の雑音的存在となってしまう．そこでクラスタリングを低頻度単語と高頻度単語とで別々に行なった．これにより，クラスタリングにおける雑音を除去し，低頻度単語の少ない情報を有効に使用する事が可能となる．

### 2．2．3　しきい値の定め方

上で求めた距離行列を用いてクラスタリングを行なうが，問題となるのがクラスタリングの際のしきい値である．そこで，このしきい値を定める目安として，品詞分散と距離歪の2つの尺度を用いる．

品詞分散$E_c$は，クラスタ内の単語の品詞的なまとまりを表す値で次のように定義する．

単語品詞ベクトル $H_T$

$$H_T = (x_1, x_2, \ldots, x_7) \qquad (2)$$

$x_i$：7つの品詞カテゴリー（名詞，形容詞，連体詞，代名詞，副詞，動詞，接続詞）に対応し，単語の品詞に該当する要素が1で他は0となる．

1つのクラスタ内の品詞分散$E_{cc}$は

$$E_{cc} = \sum_{T=1}^{N_{CL}} | H_{CL} - H_T |^2 \qquad (3)$$

但し

$$H_{CL} = \sum_{T=1}^{N_{CL}} H_T / N_{CL} \qquad (4)$$

$$|H_{CL} - H_T|^2 = \sum_{i=1}^{7} (x_i - x_i)^2 \qquad (5)$$

$N_{CL}$: クラスタ内に含まれる単語数

全クラスタに対して$E_{cc}$を求め，その和を求める事により品詞分散$E_c$が得られる．

$$E_C = \sum_{C=1}^{Q} E_{cc} \qquad (6)$$

距離歪$E_d$は，クラスタリングを行なった後に同一のクラスタに含まれた単語の距離は0と見なされ，単語間距離情報との誤差が生じるが，この誤差は元の情報の保存の指標とも考えられる．

1つのクラスタ内の距離歪$E_{dc}$は

$$E_{dc} = \sum_{i=1}^{N_{CL}-1} d_{ij} \qquad (7)$$

$d_{ij}$: 単語iがこのクラスタに含まれた時の距離

全クラスタに対して$E_{dc}$を求め，その和を求める事により距離歪$E_d$が得られる．

$$E_d = \sum_{C=1}^{Q} E_{dc} \qquad (8)$$

ここでクラスタリングのしきい値をパラメータとして，ある単語距離行列に対して，クラスタ数，品詞分散，距離歪の関係をグラフに表わした例をFig．4に示す．

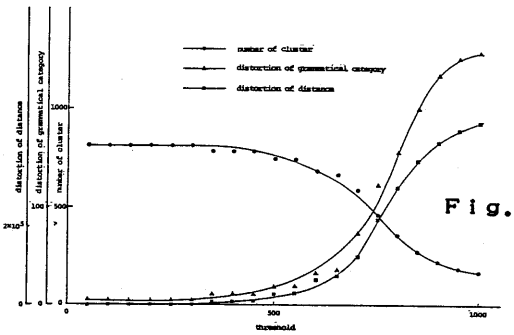



Fig. 4

Fig．4を参照する事により，クラスタリングのしきい値を，品詞分散，距離歪の急激な変化が現れる以前の約0．3付近に定めればよいと考えられる．

2．2．4　係り単語のクラスタリング

Fig．3のクラスタリング1では係り単語を先に分類している．これは，係り単語数が受け単語数より多く，又受け単語は同時に多くの単語から係られ，係り受けの特徴が現われにくいという事を考えたためである．ここで用いるアルゴリズムは，時定のクラスタに多数の単語が集中することを避けるためCentroid法[6]とした．

2．2．5　受け単語のクラスタリング

Fig．3のクラスタリング2では受け単語を分類する．前に述べたように，受け単語は，多くの意味合いで係られるために品詞分散を調べると，非常に高い値と成りやすい．そこでクラスタリングの際の時と同じ意味でCentroid法とした．

2．3　階層多重クラスタ辞書の構成

日本語の係り受けは非常に複雑で，曖昧な構造を持っている．この様な構造を捕らえるには1回のクラスタリングでは不十分だと思われる．そこで前に示した係りクラスタをもう一度分類し，しかも1つの下位クラスタは複数の上位クラスタに含まれるような階層多重クラスタを考えることにより，係り受け構造をより大きく捕らえることが可能となる．

2．3．1　階層多重クラスタの構成

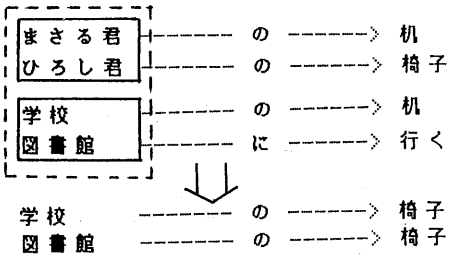
Fig3のクラスタリング3では，先のクラスタリングで構成した係りクラスタを再度分類する．

その方法としては，同じ受けクラスタに，同一の付属語を介して係っているクラスタは1つの上位クラスタに含まれるようにした．これは同じ係り受けが存在する以上，そこには広い意味でなんらかの共通点が存在するはずであるという仮定に

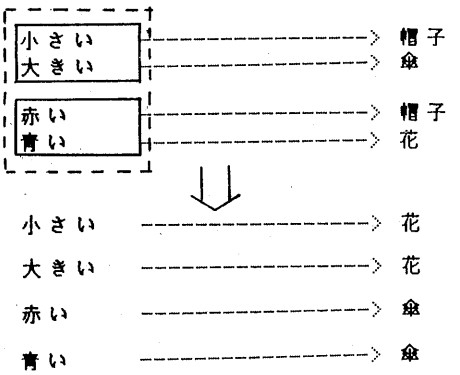
よるものである．
下位クラスタは，複数の上位クラスタに属するために，どの上位クラスタとより密接な関係があるかが順位ずけされている．

例をＦｉｇ．５に示す．点線で囲まれているのが，階層多重クラスタである．

例 1




例 2




Ｆｉｇ．５

２．３．２　階層多重クラスタの利用

クラスタを階層的にすることにより，曖昧度をそれほど増さずに，係り受けの抽象化が行なえる．つまり，ある係り受け情報を検索する場合，まず下位クラスタレベルで検索をし，その情報がそこに存在しなかった時のみ上位クラスタを検索すればよいわけで，しきい値を大きくして，大きなクラスタを構成するのにくらべて，かなり曖昧度が押さえられる．

３．　文解析実験による辞書の評価

３．１　文解析システムＥＳＳＡＹ

システムＥＳＳＡＹは，辞書駆動型のシステムであり，２．に述べた手法により構成される辞書データベースを，その知識源として用いる．システム中に手続的な形で組み込まれている知識は，日本語文の係り受けに関する原則が中心であり，文法的あるいは意味的な知識は考慮されていない．そのため２．で述べた辞書を用いた文解析実験により辞書データベースの評価を行うことが可能となる．この評価結果は，辞書構成の手法にフィードバックされる．２．において，低頻度単語と高頻度単語の処理を別けて考えるようにしたのはその例である．

３．１．１　システムの概要

システムは，文節分かち書きの形で入力された文の係り受け構造を出力する．係り受け解析のさい，基本となるのは，２文節間の係り受け関係〔５〕であり，辞書データベースを参照しつつ組み合わせ処理により係り受け構造を構成する．構成される係り受け構造は順位付け評価され出力される．処理の流れを，Ｆｉｇ．６　に示す．

３．１．２　辞書データベースの利用

（Ａ）　語彙解析：形態素テーブル，付属語テーブルを参照し，入力文の各文節を自立語部と付属語部に切り分ける．このとき，自立語付属語に対する知識（帰属クラスタ等）も検索され，また同型異義語の存在も検出される．

（Ｂ）係り受け可能性の検索：各文節に対して，係り得る文節を，クラスタ間係り受け辞書の検索によりすべて求める．このとき，単語辞書，付属語辞書中の統計量を基にＢａｙｅｓの定理を用いて，係り受けの機能を表わす係り受けラベルの推定を行う．また係り受けの良さをローカルに評価し，係り先候補をソートする．単語クラスタ情報のみでは係り先

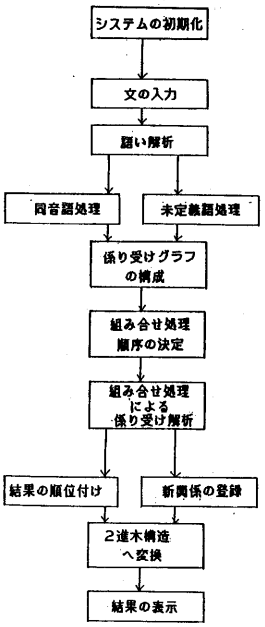



Fig. 6 ESSAY における処理の流れ

候補文節が求められない場合，階層多重クラスタ情報を用いる．

3．1．3　解析結果の順位付け

係り受け可能性をリンクとし，文節をノードとするグラフ構造から解析結果を表わす係り受け構造を組み合わせ処理により抽出する．この組み合わせ処理を行う順序を，係り受けリンクに対するローカルな評価値を基に制御する．この段階で，解析数は評価しうるので，この数が設定した限度数より多い場合は，刈り取り処理を行う．本処理により正しい解析が落ちる率は，限度数を５０とした場合４％程度であった．組み合わせ処理により構成された係り受け構造の良さは，次の評価関数により求められ，評価値に従い順位付けを行う．式中，A，B，Cは，０～１の値を取るパラメータであり，その最適近似値は，正しい結果の平均順位，正しい結果が１位に出力されない率を指標とした予備実験によって，（A，B，C）＝（０．１，０．９，０．９）とした．２進木構造に変換された出力例を，Ｆｉｇ．７に示す．

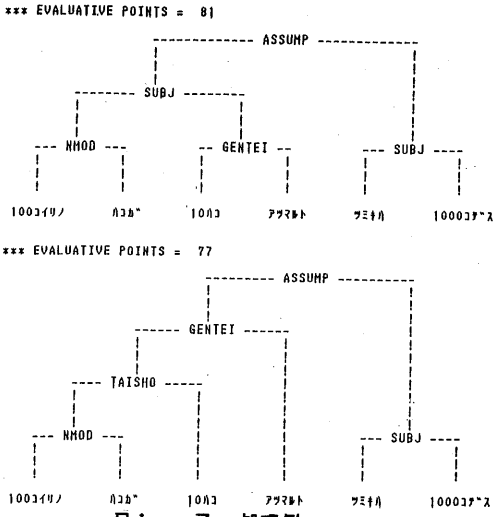



Fig. 7　処理例

$$E = \frac{\sum_{k=2}^{n} U(k)\ C(k)}{\sum_{k=2}^{n} U(k)} \quad (9)$$

$$U(k) = \begin{cases} 1 & \cdots \text{k文節が受け文節} \\ 0 & \cdots \text{それ以外} \end{cases} \quad (10)$$

$$C(K) = B\left(\sum_{j=1}^{k-1} KU(j,k)\ (C\ g(j,k) + (1 - C)\ E_{jk})\right) / \sum_{j=1}^{k-1} KU(j,k) + (1 - B)\ L_a(k) \quad (11)$$

$$KU(j,k) = \begin{cases} 1 & \cdots \text{j文節→k文節} \\ 0 & \cdots \text{それ以外} \end{cases} \quad (12)$$

$$g(j,k) = \begin{cases} 1 & \cdots \text{品詞レベルでマッチ} \\ 0 & \cdots \text{それ以外} \end{cases} \quad (13)$$

$$E_{jk} = A\ P_{jk} + (1 - A)\ W_{jk} \quad (14)$$

$P_{jk}$：ラベル推定の尤度
$W_{jk}$：辞書中の尤度
$E_{jk}$：ローカルな評価値

$$L_a(k) = \frac{\text{文中で埋められたスロット数}}{\text{k文節中自立語の受けスロット数}} \quad (15)$$

3．2　文解析実験と評価

初期辞書の構成に用いない文５００文に対し実験を行った結果をＴａｂｌｅ．１に示す．表中，カバーリンク数とは，処理可能な係り受け単語のペア数である．クラスタリングを行うことによる抽象化の効果が現れているが，得られる解析率は，十分とは言えない．そこで，失敗の場合を考察したところ，係り先を求められない文節をふくむ場合，その文節数の

割合は３０％程度であり，学習処理の可能性と必要性が確認された．

Ｔａｂｌｅ．１

| | 単語レベル | クラスタレベル | 階層使用 |
|---|---|---|---|
| 係り | | | |
| クラスタ数 | １５２９ | ８３９ | …… |
| 受け | | | |
| クラスタ数 | ９６０ | １２１５ | …… |
| 係り受け | | | |
| リンク数 | ６１６２ | ４２０２ | …… |
| カバー | | | |
| リンク数 | ６１６２ | １０６８１９ | …… |
| 品詞分散 | ０ | ３５．４ | …… |
| 文解析率(%) | ３２．５ | ４６．７ | ５４．２ |

## ４．　学習情報の獲得と構造化

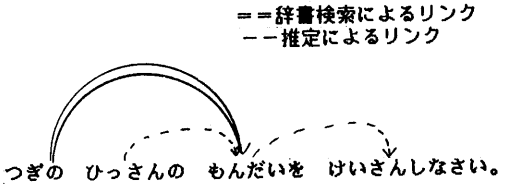
学習処理は，学習情報を文解析処理を通して獲得するフェーズと，獲得した学習情報を既存の構造中に構造化するフェーズの２フェーズにより行われる．学習情報としては，現在の辞書中に存在しない語に関する情報と，現在の辞書記述では処理できない係り受け関係に関する情報の２種類が考えられるが，前者はユーザとの対話により単語情報を得，新たに単語クラスタを割り付けることにより行われる．以下では後者について扱う．これらの処理に対する支援機能はシステムＥＳＳＡＹ中に用意されている．

### ４．１　文解析からの学習情報の獲得

文解析におけるネガティブな状況（解析の失敗，正しい結果が上位に出力されない）は，使用中の辞書の性能が十分でないことを示すものであるが，逆に考えれば，そのような状況を生じさせた原因に関する情報を抽出し，辞書の内容にフィードバックすれば，同じ原因による失敗は以後生じなくすることが可能となる．このような獲得の処理は，文解析をオンラインで行う場合と，オフラインで行う場合とでは現在のところ異なっている．すなわち，オフラインの場合，正しい結果を入力文ファイル中に付加しておき，出力される解析結果ファイルとの突き合わせ処理により行うが，オンラインの場合，システム側の推定とユーザの対話により行うこととしている．

#### ４．１．１　係り受け構造の推定

入力文が正しい文であり，辞書検索のより得られている構造が正しいとの仮定の下では，推定により係り受け構造が決定できる場合が存在する．簡単な例をＦｉｇ．８に示す．この処理は，係り受け構造の性質に基づいてなされるが，一般にはこの場合のように一意に構造を定めることは不可能である．そこでＦｉｇ．９に示すようなアルゴリズムに基づきシステムの推定とユーザの応答により正しい構造を決定し，学習情報の獲得を行う．




Ｆｉｇ．８

文節は文頭側から文末側へ昇順に，１，．．．，ｎと番号付けされているとする．図中で，未定文節とは辞書検索により係り先が見出されなかった文節をいい，限界文節とは，未定文節がこれを越えては係り得ない文節をいう．限界文節及び係り先候補は，非交差条件を考慮して求められる．また，ＥＶＡＬＣ（ｉ，ｊ）とは，文節ｉ～文節ｊまでの部分構造に対する評価関数であり，現在のところ次のように定義している．

$$EVALC(i,j) = \sum_{k=i}^{j-1} |d(k) - k - f(k)| \qquad (16)$$

ここで，ｆ（ｋ）とは文節ｋの付属語が係り文節に存在した場合の，係り先までの隔たりの平均値であり，付属語テーブルに記述されている．また，ｄ（ｋ）は文節ｋの係り先文節の番号を表わす．

```
for i := n - 1 to 1
  if 文節 i は未定文節 then
    begin
      LM := 限界文節;
      if LM - i = 1  then
        (i -> LM) を登録
      else begin
        NC := 係り先候補数;
        C_LIST := 係り先候補リスト;
        if NC = 1  then
          (i -> NC) を登録
        else begin
          for j := 1 to NC;
            VALUE(j) := EVAL(i, C_LIST(j))
          C_LIST を VALUE に従いソート;
          for k := 1 to NC;
            if 同い合せ(i -> C_LIST(k)) = OK
              then begin
                (i -> C_LIST(k)) を登録;
                break;
              end
        end
      end
    end
```

Fig. 9 推定アルゴリズム

### 4．1．2 係り受け尤度の適応化

本値は（14）式中のWであり，解析結果の順位付けに際して基本となる量である．正しい結果が1位に出力されなかった場合，本値に対し，適応化処理を行う．

### 4．2 学習情報の構造化

4．1に述べた処理により獲得される新規の係り受けに関する情報は，単にクラスタ間の係り受け関係として既存の構造に新しく付加するよりも，構造化したほうが合理的であり，獲得した学習情報を構造化して保持することも可能となる．

本アプローチにおいては，単語はクラスタ化されるので，構造化処理は単語クラスタ形状を再構成する処理であると言える．しかしながら，現在の初期辞書の構成手法は2．に述べたように，かなり複雑な処理によっているので，文解析された文をすべて原テキストデータに付加し，すべての処理を再度行うことは得策とは言えない．

### 4．2．1 単語クラスタ形状の再構成

そこで，1次的な知識構造である単語クラスタ形状の再構成を，Fig．10に示す処理により行うこととした．

ここで，特徴的なことは一時的に上位レベルの情報（一時的階層クラスタとよぶ）を用いることである．

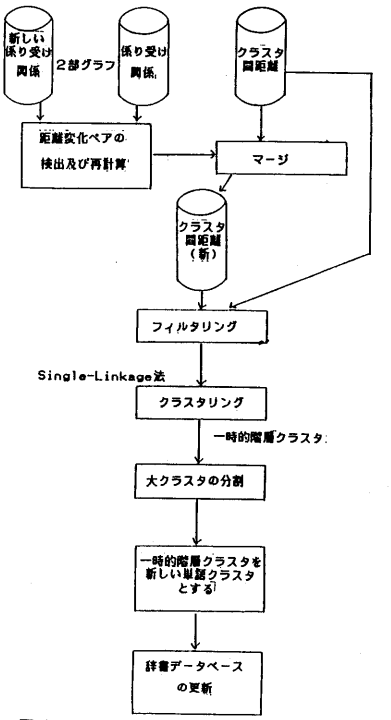



Fig. 10 単語クラスタ形状の再構成

（A）距離変化ペアの検出と再計算：クラスタi，j間の距離を次式で定義する．

$$D_{ij} = 1 - \frac{\sum_{r} (w(i \to r) + w(j \to r))}{\sum_{p} w(i \to p) + \sum_{q} w(j \to q)}$$

w(*)はリンクの頻度 (17)

このように，距離はクラスタ間係り受けの2部グラフ上で定義されるため，新しいリンクの付加により距離の変化するペアは容易に検出できる．よって，これらのペアについてのみ，距離の再計算を行えばよい．

（B）フィルタリング：クラスタリングに対して入力となるのは，クラスタ間の距離行列である（新，旧をそれぞれD，D'とする）．このとき，新旧の状況の差に関する情報を用いるため次の変化率行列Rを定義し，

$$R = (R_{ij}) = \left( \frac{D'_{ij}}{D_{ij}} \right) \qquad (18)$$

このＲｉｊが閾値条件を満たさないペアは，クラスタリングの対象としないという前処理を行った．

（Ｃ）クラスタリング：用いたアルゴリズムは，Ｓｉｎｇｌｅ－Ｌｉｎｋａｇｅ法〔６〕である．（Ｂ）の前処理により，本アルゴリズム特有のｃｈａｉｎｉｎｇ－ｅｆｆｅｃｔはある程度さけることができた．

（Ｄ）大クラスタの分割：（Ｃ）により構成される一時的階層クラスタには，サイズの大きいものも含まれるため，要素のサイズを考慮し，幾つかの核クラスタを選出し，細分割を行った．

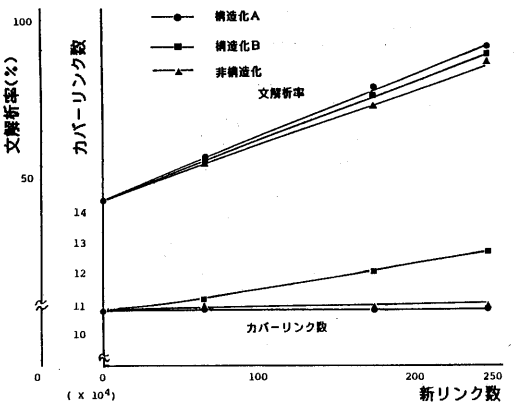
４．２．２　階層クラスタ形状の再構成

学習情報を考慮した階層多重クラスタ形状の更新は，単語クラスタ間の係り受けリンクに新規のものを付加し，Ｆｉｇ．３のクラスタリング３の処理により行う．本処理は，簡易に行えることがその特長である．

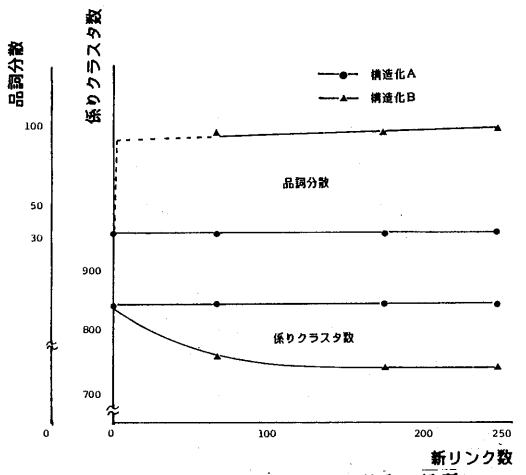
なお，４．２に述べた処理は係りクラスタについてのみ行う．

５．学習実験と評価

学習処理により再構成された辞書データベースの性能と性質の，学習情報の量に対する変化を調べた．性能については，カバーリング数と文解析率を指標とした．結果をＦｉｇ．１１に示す．図中，構造化Ａとはすべての処理を再度行うものであり，構造化Ｂとは４．２に述べた処理によるものである．また非構造化とは，単にリンクを付加したものである．構造化による抽象化の効果が表れていることがわかる．性質については，クラスタ数と品詞分散を指標とした．結果をＦｉｇ．１２に示す．構造化Ｂの場合，学習情報の加え始めにおいてｊｕｍｐが見られるが，双方の量ともほぼ一定の傾向を示すことがわかり，新しい単語を含まないとする条件下では，新しいリンクを考慮しても大きな変化は生じないことがわかる．




Ｆｉｇ．１１　学習情報の量と辞書の性能




Ｆｉｇ．１２　学習情報の量と辞書の性質

６．　まとめ

本アプローチに基づく，知識構造の構成及び成長方法について有効性が確かめられた．今後は，４．１．１の考え方を推進し，自律学習の可能性をさぐるとともに階層的な情報をより有効に利用する手法を，解析システムに組み込んでいく予定である．

《参考文献》
[1] 比井：表現と利用　信学技報 AL-80-27 (1980)
[2] Shirai,K., et.al: Japanese Sentence Analysis System ESSAY, COLING 82 (1982)
[3][4] 白井，林，平田：情報処理学会第２４回全国大会 3K-4, 5 (1982)
[5] 吉田：２文節間の係り受けを基礎とした日本語の構文解析 信学論D Vol.55-D No.4 (1972)
[6] Hartigan,J.A.:Clustering Algorithms John Wiley & Sons (1975)