漢字複合語の確率的構造解析†*

西野哲朗†† 藤崎哲之助†††

従来の研究から，漢字複合語の自動分割についてはかなりの成果が得られているが，複合語がただ分割されているだけでは，その後の段階で十分な意味処理を行うことは難しい．そこで我々は，複合語の構造を決定することが，より高度な意味処理を実現する上で重要であると考え，漢字複合語の確率的構造解析の研究を進めている．本稿では，その手法と実験結果について述べる．我々の手法においては，漢字複合語の構造は，文脈自由文法によってモデル化される．その際，漢字複合語を構成する各漢字短単位語に対して，自然言語文における単語の品詞に対応する概念を定義する必要が生じる．この目的のために，各短単位語の複合語内における共起関係の頻度情報を用いて，それらのクラスタリングを行い，得られたクラスタを品詞に対応する概念として用いた．そして次に，上記の文脈自由文法の曖昧性を除去するために，実際のデータにより確率付けされたその文脈自由文法によって漢字複合語を構造解析して，各複合語に対する最も自然な解析木を得ることを目指した．また，本構造解析手法の一つの応用として，かな漢字変換エラーの自動検出の実験を行ったので，その結果も併せて報告する．

1. まえがき

漢字複合語の自動分割は，機械翻訳，音声合成等で必要となる日本語処理の基礎技術であり，すでにかなりの成果が得られている[1),2)]．しかし本来，漢字複合語の解析は，意味処理の前段階としての，形態素解析の一部分と考えられるべきものである．したがって，漢字複合語がただ分割されているだけでは，その後の意味処理が十分に行えない場合が多い．そのような場合に最大の問題であると考えられるのは，漢字複合語の構造が決定されていないことである．例えば機械翻訳の場合，複合語の構造が決定されていなければ，複合語の意味の違いによる訳し分けや，文生成時における複合語への助詞の挿入等において大きな困難が生じる．

このような現状において我々は，漢字複合語の構造解析がより高度な日本語処理への鍵になると考え，その自動化についての研究を行った[3),4)]．従来，漢字複合語の解析は，国立国語研究所の語彙調査等で人手により行われていただけで[5)~8)]，自動化の試みは今回の研究が初めてである．

漢字複合語の構造解析を行うにあたっては，さまざまな方法が考えられるが，我々は，自然言語文の構文解析のアナロジーとして次のようなアプローチを取った．すなわち，漢字複合語の構造を文脈自由文法によってモデル化し，一般の構文解析アルゴリズムにより，その構造解析を行った．この手法を取る場合，以下のような二つの大きな問題をまず解決しなければならない．

1. 漢字複合語を構成する各漢字短単位語に対し，自然言語文における単語の品詞に対応する概念を定義する．
2. 漢字複合語の構造は文の構造より単純なので，普通に構造解析すると，複数の解析木を得る語が大部分を占める．このような文脈自由文法の曖昧性を除去する．

漢字複合語の構造解析が，それら複合語の分割結果をもとに行われるのは，自然な処理の流れであろう．したがって上記項目1の問題は，漢字複合語を構成する短単位語を，その構文的機能により分類する問題に帰着される．

漢字複合語解析の自動化が今までに行われなかった理由の一つとして，漢字複合語を構成する短単位語には，例えば品詞のような，構造解析を行う上での手がかりとなる，構造に関する確立された情報がなかったことがあげられる．また，各短単位語に人手で構文情報等を付与するのは，きわめて困難である．従来，漢字短単位語に対しては，分類語彙表のような意味分類は行われてきたが，各短単位語の漢字複合語内における構造上の位置の分布に基づいた分類は，行われてこなかった．また，解析の際に現れる短単位語の種類は，それらが構成する漢字複合語が出現した文献の分

† A Stochastic Parsing of Kanji Compound Words by Tetsuro Nishino (Department of Information Sciences, College of Science and Engineering, Tokyo Denki University) and Tetsunosuke Fujisaki (Tokyo Research Laboratory, IBM Japan Ltd.).
†† 東京電機大学理工学部情報科学科
††† 日本アイ・ビー・エム(株)東京基礎研究所
* 本研究のほとんどは，第一著者が日本アイ・ビー・エム(株)，東京基礎研究所在職中に行われた．

野に大きく依存する．我々が解析の対象とした科学技術文献の場合，既存の分類には現れていない短単位語が多数を占めていた．

そこで我々は，分野が一つ特定されたときに，そこに出現する短単位語をその構文的機能により分類するために，同一複合語内におけるそれらの共起関係の頻度情報を用いて，それら短単位語をクラスタリングした．本稿第2章では，その漢字短単位語のクラスタリングについて述べる．

このようにクラスタリングを行うと，一つの複合語内に出現したある短単位語をその語と同じクラスタに属する別の語で置き換えたとしても，その結果得られる複合語は，意味は変わるが依然として適格な複合語であると期待できる．さらに，そのような交換可能な語から成るクラスタは，意味的に関連がある語によって形成されていると推測できる．このように短単位語間の共起関係という複合語内での位置に関する統計情報から，短単位語の意味情報を抽出できるということを示すのも本研究の目的の一つである．

上記項目2の問題は，自然言語文の構文解析においても重要な問題であり，すでに曖昧性除去の手法がいくつか提案されている．我々は，そのような手法のうち，英語文の構文解析における曖昧性除去ですでにその有効性を実証した，確率的構文解析の手法[9),10)]を漢字複合語の解析に適用した．そこでは，実際のデータにより確率付けされた文脈自由文法によって漢字複合語を構造解析し，各複合語に対する最も自然な解析木を得ることを目指した．そのことについては，第3章で述べる．

上でも述べたように，漢字複合語の構造解析は，多くの有用な応用の基礎になるものと期待できる．そこで我々は，この確率的構造解析手法の応用における有効性を探る意味から，かな漢字変換エラーの自動検出の実験も併せて行った．その結果については，第4章で報告する．また第5章で，現バージョンの問題点についての考察を行う．

2. 漢字短単位語のクラスタリング

2.1 入力データの形式

以下では，三文字以上の漢字列で後述の条件を満たすものを，漢字複合語と定義する．漢字複合語を構成する一文字および二文字の短単位語を，接頭辞，接尾辞および二文字語と呼んで区別する．例えば，「超大型計算機」という漢字複合語は，「超」という接頭辞，「大型」，「計算」という二つの二文字語および，「機」という接尾辞から構成されている．本論文で対象とする漢字複合語は，次の条件を満たすものとする．

1. 二文字語を必ず一つは含む．
2. 接頭辞の直後に接尾辞が現れることはない．

この漢字複合語の構造は，基本的に武田らのモデル[1)]に従っている．

漢字短単位語クラスタリングの処理の対象となる入力データは，短単位語に分割され，かつそれぞれの短単位語が二文字語，接頭辞および接尾辞のうちのどれであるかが，決定されている漢字複合語である．入力データの形式を図1に示す．この部分の複合語分割処理は，すでに自動化されている[1)]．

2.2 手続き

短単位語のクラスタリングの手続きを，図2に示す．

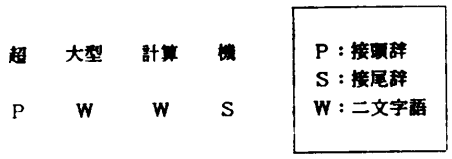


図1 入力データの形式
Fig. 1 Input data format.

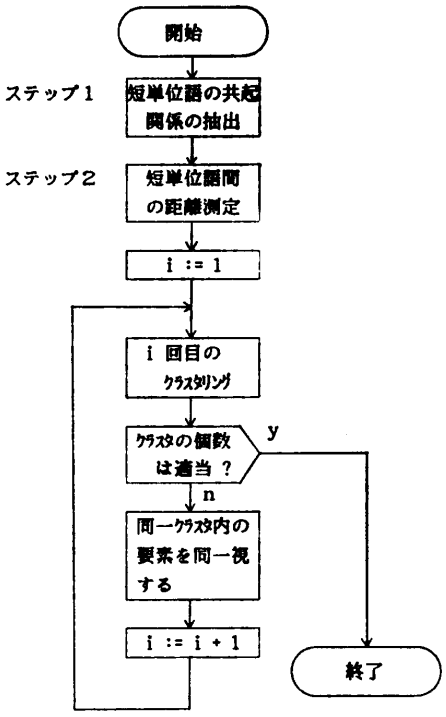


図2 短単位語のクラスタリングの手続き
Fig. 2 Procedure of Kanji primitive word clustering.

2.2.1 漢字短単位語間の共起関係の抽出

ステップ1では，図1の形式の多数の複合語データから，短単位語の共起関係を抽出する．ここで，二つの短単位語 X と Y が共起するとは，X と Y が同一複合語内に現れており，かつ X と Y が係り受けをしている可能性があることと定める．このような共起関係にある複合語の組を抽出するためには，同一複合語内に現れた短単位語のすべての組から，係り受け関係になりえない短単位語の組を除外しなければならない．そのような除外すべき場合としては，次の二つの場合が考えられる．

（a） 接頭辞は，自分の左側の短単位語とは係り受けをしない．

（b） 接尾辞が右隣にある短単位語（二文字語または接尾辞）は，その右隣の接尾辞を越えてさらに右側にある短単位語とは係り受けをしない．

これら二つの場合に対応する共通関係抽出の規則は，次のようになる：X を短単位語とし，複合語内で X の右隣に現れる短単位語を R とする．

（1） R が接頭辞または二文字語のときは，その複合語内で X の右側に現れる接頭辞を除くすべての短単位語（二文字語または接尾辞） Y との関係 (X, Y) を抽出する．

（2） R が接尾辞のときには，関係 (X, R) のみを抽出する．（この場合，漢字複合語モデルの構造から，接頭辞の右隣に接尾辞が現れることはないので，X は接頭辞ではありえない．）

例えば「逆多項式変換」という複合語からは，**図3**に示した五つの共起関係が抽出される．この規則に従ってステップ1では，入力データに出現した各短単位語が，どんな短単位語と何回同一複合語内に共起したかに関する頻度情報をとる．ステップ1の出力例を**図4**に示す．図4では，二文字語「石油」と「石炭」が，JICST 文献抄録に含まれていた 31,900 語の漢字複合語内で，それぞれどんな短単位語と何回共起したかが示されている．

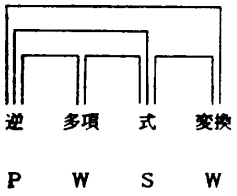


図3 漢字複合語から抽出される共起関係
Fig. 3 Co-occurrence relation extracted from a Kanji compound word.

2.2.2 漢字短単位語間の類似度の定義

ステップ2では，ステップ1の結果をもとに，任意の二つの短単位語間の距離を計算する．直観的に言えば，図4の例の場合，四角で囲まれた二つの部分 P と Q の類似性から，「石油」と「石炭」の類似度が決定される．形式的な定義を以下に述べる．

解析の対象となる領域に出現した短単位語の異なり数を m とし，各短単位語は，それら m 語のうちの何番目の語であるかが指定されているものとする．短単位語 A の共起ベクトルとは，その第 i 成分 $(1 \leqq i \leqq m)$ が，A がその領域の i 番目の短単位語と，同一複合語内に共起した回数であるような m 次元ベクトルのこととする．

二つの短単位語 A と B の共起ベクトルをそれぞれ a, b とするとき，A, B 間の類似度 $d(A, B)$ をベクトル a と b の内積の値を用いて次のように定義する．

$$d(A, B)=\frac{a \cdot b}{|a| \times |b|} \qquad (1)$$

ここに，$|a|$ はベクトル a の長さである．式(1)で定義される類似度が大きい二つの短単位語ほど，それら二語の共起ベクトルが近い方向を向いているので，複合語を形成する場合の使われ方が類似していることになる．定義からわかるように，短単位語間の類似度は0以上1以下の実数値をとる．

具体例として，2.3.1 節，図4の場合について考える．d(石油，石炭) を求めるには，次のように計算

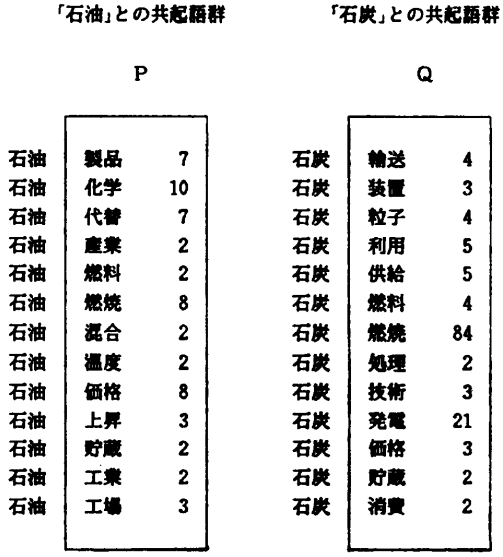


図4 共起に関する頻度情報の例
Fig. 4 Examples of the frequency of co-occurrence.

を行えばよい．すなわち，図4からわかるように，石油と石炭の双方と同一複合語内で共起した語は，燃料，燃焼，価格，貯蔵の四語のみである．ゆえに，これら四語以外の語に対応した成分は，二つの共起ベクトルのうちの少なくとも一方において0である．したがって，石油と石炭の共起ベクトルの内積を計算する際に，この四語以外に対応した成分どうしの積をとると，その値は0となり内積の値に貢献しない．また，図4に現れない語とは，石油も石炭も共起しなかったことを考え合わせると，図4の情報だけから，d(石油，石炭) を計算できる．図4の場合，四角 P と Q の中に現れる数の二乗和はそれぞれ 368，7634 であるから，d(石油，石炭) の値は，

$$\frac{2\times4+8\times84+8\times3+2\times2}{\sqrt{368}\times\sqrt{7634}}=0.422$$

となる．

2.2.3 クラスタリングの方法

2.2.2 節で求まった類似度を用い，最短距離法 (nearest neighbour method) によってクラスタリングを行う．二つの短単位語は，類似度が大きいほど距離が近いと定義する．各短単位語は，ただ一つのクラスタに割り当てられる．最短距離法においては，クラスタ間の距離として，それぞれのクラスタに含まれている最も近い要素間の距離を採用する．その場合クラスタどうしが融合すると，そのまわりの部分が融合した部分に近づくことになる．こうしてクラスタの融合によって空間が濃縮され，点と点との間に鎖が生じて鎖状のクラスタができる．このような濃縮という特徴から，最短距離法の分類感度は低いと考えられている[11]．

そこで我々は，重心法 (centroid method) によるクラスタリングも併せて行い，それぞれの手法によるクラスタリング結果を比較した．その結果，今回の我々の実験に関しては，類似度の高い語が次々と同じクラスタに取り込まれて行く最短距離法の方が，人間の直観に近い分類ができると判断した．

2.3 実験結果

JICST の文献抄録から抽出された，長さ三以上の複合語のべ 31,900 語をもとに，短単位語のクラスタリングを行った．クラスタリングは二文字語（約 1,000 語），接頭辞（約 200 語）および，接尾辞（約 400 語）の各場合について別々に行い，それぞれを 28 個，8 個および，10個のクラスタに分類した．このようにクラスタリングを短単位語の形態別に行ったのは，例えば接尾辞はどのような語とも共起してしまい，二文字語とはかなり異なった性質を示したためである．

得られた接頭辞クラスタと接尾辞クラスタの例を図5に示す．図5(a)の接頭辞クラスタは主に数詞から，(b)の接尾辞クラスタは主に体の部分に関する語から形成されている．

形成されたクラスタに直観に反する語が入る現象は，その語に関する情報が非常に少ないときによく見うけられた．例えば，接頭辞クラスタ P01 に「英」が含まれる理由は，「英」に関する頻度情報が少なか

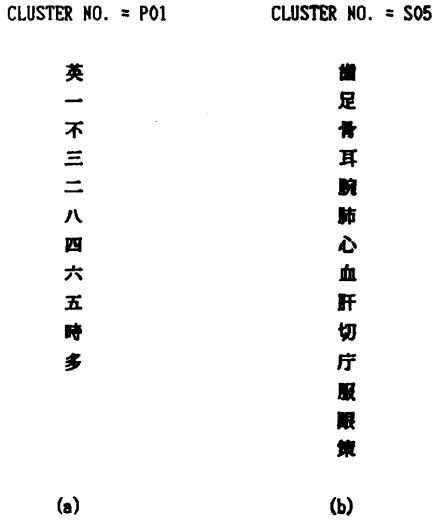


図5 得られたクラスタの例
Fig. 5 Examples of the obtained clusters.

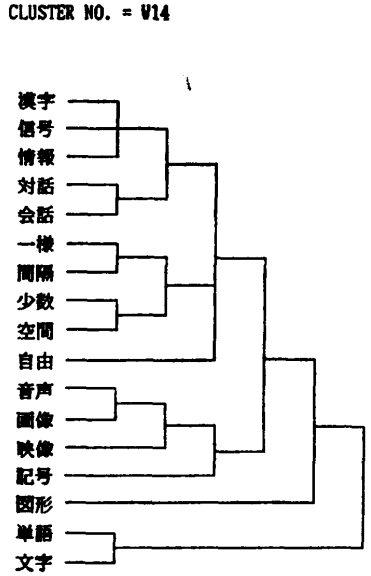


図6 クラスタの階層構造
Fig. 6 Hierarchical structure of a cluster.

ったために，たまたま現れた「英単語」が，「一単語」，「二単語」等の複合語と類似するという現象が生じたためである．

得られたクラスタは，図2のような繰返し手続きによって形成されたものなので，実際には階層構造を持っている．その様子を，二文字語クラスタの場合について**図6**に示す．このクラスタは，主にコミュニケーションに関する語から形成されている．図6から，「対話」と「会話」，「単語」と「文字」等の語がまず最初にクラスタを成し，その後次第に，より大きなクラスタが形成されていく様子がわかる．

規則の型			規則の数
Z	-> C[v]		38
C[v2]	-> C[v1]	C[v2]	1,444
C[v]	-> x	C[v]	304
C[u]	-> C[v]	u	380
C[v]	-> y	C[v]	1,064
C[y]	-> C[v]	y	1,064
C[y]	-> x	y	224
C[u]	-> y	u	280
C[y2]	-> y1	y2	784
		合計	5,582

x : 接頭辞クラスター {P01,P02,...,P08}
y,y1,y2 : 二文字語クラスター {W01,W02,...,W28}
u : 接尾辞クラスター {S01,S02,...,S10}
v,v1,v2 : {W01,...,W28,S01,...,S10}

図8 複合語解析のための文脈自由文法の規則
Fig. 8 Rules of a context-free grammar used for the Kanji compound word analysis.

3. 漢字複合語の確率的構造解析

3.1 漢字複合語の構造のモデル化

3.1.1 構造に関する仮定

我々は，漢字複合語の構造を，**図7**に示すような解析木で表現する．解析木が，二分木になっていることに注意されたい．解析木をすべて二分木で表現することは，我々が事前の調査から設けた漢字複合語の構造に関する仮定である．図7の解析木に現れる非終端記号は，次のような二つの大きな特徴を持っている．

1) 各短単位語は，一つのクラスタに属しているが，そのクラスタの名前が，非終端記号に反映されている．

図7において，「大型」という二文字語が W17 という非終端から導出されているのは，「大型」が二文字語の 17 番目のクラスタに属していることを示している．

2) 複合語のカテゴリーには，それを構成する後方の短単位語または複合語のカテゴリーを反映させる．

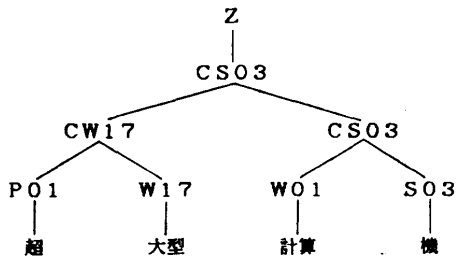


図7 漢字複合語の解析木
Fig. 7 A parsing tree of a Kanji compound word.

図7において，「超大型」が CW17 という非終端に支配されているのは，「超大型」という複合語の属するカテゴリーは，「大型」という短単位語の属するカテゴリーによって決定されることを表している．すなわち，「超大型」という概念は，「大型」という概念の特殊な場合であると考える．これも，我々が事前の調査から設けた仮定である．

3.1.2 文脈自由文法によるモデル化

3.1.1 節の仮定により，任意の漢字複合語は，Chomsky Normal Form の文脈自由文法で構造解析することができる．その文法規則の一部を**図8**に示す．今回の我々の実験では，二文字語，接頭辞および接尾辞は，それぞれ 28 個，8個および 10 個のクラスタに分割されていたので (2.3節の実験結果を用いた)，各型の規則は図8に示した個数となった．

図8において，x は接頭辞クラスタの名前全体の集合 {P01, P02, ···, P08} の要素，y, $y1$, $y2$ は二文字語クラスタの名前全体の集合 {W01, W02, ···, W28} の要素，u は接尾辞クラスタの名前全体の集合 {S01, S02, ···, S10} の要素，そして v, $v1$, $v2$ は名前の集合 {W01, ···, W28, S01, ···, S10} の要素である．図中，例えば C[v] という記法は，文字 'C' と v が取り得る名前（文字列）の任意の連接 (CW01, CW02, ···等) を表している．

複合語解析に用いた文脈自由文法は，漢字短単位語を終端記号とし，図8の規則に現れる形の非終端記号を持つ．開始記号はZである．図8では，終端記号を導出する規則は省略した．

3.2 確率的パージングの枠組み

ここでは，漢字複合語の解析から一時はなれて，確率的構文解析の一般的枠組みについて，その概略を紹介する．詳しくは，文献[9),10)]を参照されたい．

図 9 に確率的構文解析の枠組みを図式的に示した．与えられた曖昧性のある線形文法を，例文より反復計算によって確率付けする手法として，Forward and Backward アルゴリズムが知られている[12)]．図 9 に示した枠組みでは，その Forward and Backward アルゴリズムを一般の文脈自由文法に拡張したものが用いられている．

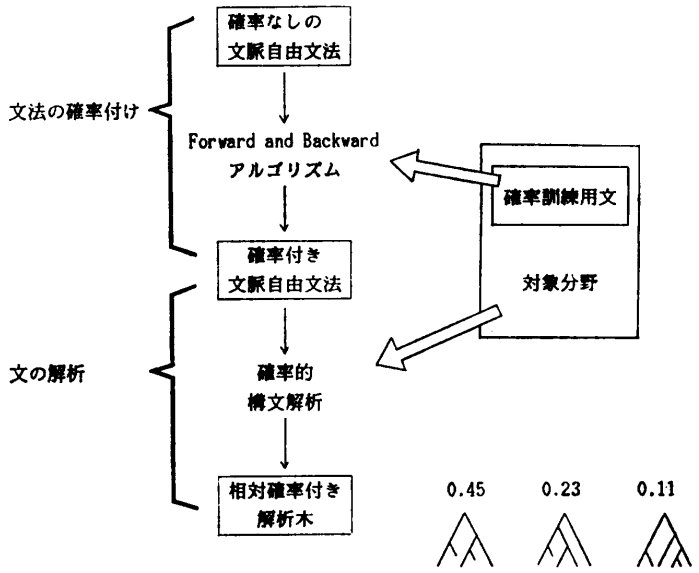


図 9 確率的構文解析の枠組み
Fig. 9 Framework of the stochastic parsing.

各規則に確率が付与されていない普通の文脈自由文法に対し，Forward and Backward アルゴリズムを用いることにより，各規則が訓練用文を生成する際に使用された相対頻度から，それら規則が対象分野の文を生成する際に使用される確率を推定することができる．そうして得られた確率付き文脈自由文法を用いて構文解析を行うと，入力文に対して複数の解析木が存在するときには，各解析木に相対確率が付与される．最大確率が付与された解析木が，入力文に対して最も自然な解析木であるとき，確率的構文解析は成功したと言われる．

3.3 実験結果

実験に用いた確率的構造解析プログラムの構成を，**図 10** に示す．実験には，前もって短単位語に分割された漢字複合語データを用いた．短単位分割の処理は，すでに自動化されていた[1)]．実験は次のようなステップで行われた．

1) 辞書引き：辞書には，各短単位語が所属するクラスタが登録されている．

2) 28,568 個の訓練用データによる文法の確率付け：文法は，図 8 に示した形式の 5,582 個の規則と，終端記号（短単位語）を導出する規則から成る．(図 9 も参照されたい．)

3) 2) の訓練用データには含まれていなかった，1,636 個のテストデータでの確率的構造解析：このテスト結果から，153 件をランダム・サンプリングし，**表 1** のような結果を得た．テストデータでの成功率 0.736 に対し，訓練用データでの成功率は 0.76 であった．テストデータについて出力された解析木の例を，**図 11** に示した．図 11 (a) で A：0.215(0.13) とあ

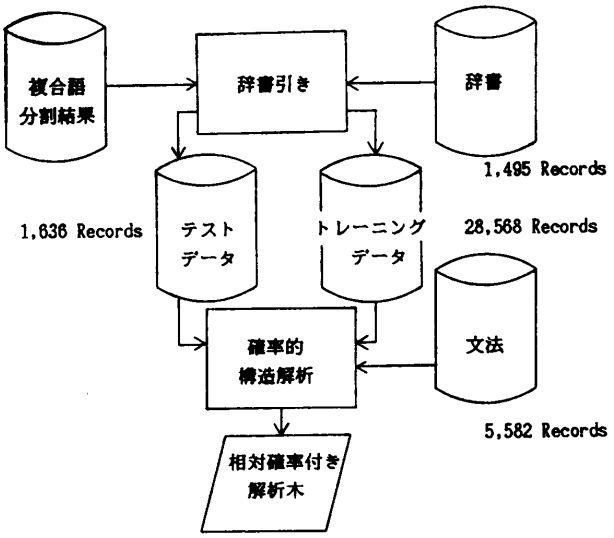


図 10 システム構成
Fig. 10 Construction of the system.

るのは，そのAの右側の二つの解析木が現れる相対確率が，訓練後に 0.215 であったことを示している．カッコ内の 0.13 は訓練前の相対確率であり，初期確率から算出された値である．図 11 (a) で最大確率を得たのは最下段の解析木で，その相対確率は

$$0.600 \times 0.711 = 0.427$$

である．またその解析木は

(((中規模) 集積) 回路)

という構造をしており，最も自然な解析木である．なお結果に対する考察は，第 5 章で述べる．

```
1** Sentence No.   83 **
 <中   規模 集積 回路 . >
 ** total ambiguity is :      5
 *:
  A: 0.215(0.13)  SE          <L0000001>
   |A: 0.807(0.65)  CW01        <L0005165>
   | |*:              P08         '中
   | |*:              CW01        <L0003835>
   | |  *:              CW01        <L0000564>
   | |    *:              W22         '規模
   | |    *:              W01         '集積
   | |  *:              W01         '回路
   |B: 0.193(0.35)  CW01        <L0005165>
   | |*:              P08         '中
   | |*:              CW01        <L0003569>
   | |  *:              W22         '規模
   | |  *:              CW01        <L0000543>
   | |    *:              W01         '集積
   | |    *:              W01         '回路
   |*:              PRD         '.
  B: 0.185(0.45)  SE          <L0000001>
   |*:              CW01        <L0001348>
   |  *:              CW22        <L0000214>
   |    *:              P08         '中
   |    *:              W22         '規模
   |  *:              CW01        <L0000543>
   |    *:              W01         '集積
   |    *:              W01         '回路
   |*:              PRD         '.
  C: 0.600(0.42)  SE          <L0000001>
   |A: 0.289(0.21)  CW01        <L0003835>
   | |*:              CW01        <L0005165>
   | |  *:              P08         '中
   | |  *:              CW01        <L0000564>
   | |    *:              W22         '規模
   | |    *:              W01         '集積
   | |*:              W01         '回路
   |B: 0.711(0.79)  CW01        <L0003835>
   | |*:              CW01        <L0003856>
   | |  *:              CW22        <L0000214>
   | |    *:              P08         '中
   | |    *:              W22         '規模
   | |  *:              W01         '集積
   | |*:              W01         '回路
   |*:              PRD         '.
```

(a)

```
1** Sentence No.   89 **
 <小   規模 電力 会社 . >
 ** total ambiguity is :      5
 *:
  A: 0.000(0.21)  SE          <L0000020>
   |A: 1.000(0.25)  CW20        <L0005184>
   | |*:              P08         '小
   | |*:              CW20        <L0004564>
   | |  *:              CW08        <L0000760>
   | |    *:              W22         '規模
   | |    *:              W08         '電力
   | |  *:              W20         '会社
   |B: 0.000(0.75)  CW20        <L0005184>
   | |*:              P08         '小
   | |*:              CW20        <L0003588>
   | |  *:              W22         '規模
   | |  *:              CW20        <L0001082>
   | |    *:              W08         '電力
   | |    *:              W20         '会社
   |*:              PRD         '.
  B: 1.000(0.71)  SE          <L0000020>
   |*:              CW20        <L0002070>
   |  *:              CW22        <L0000214>
   |    *:              P08         '小
   |    *:              W22         '規模
   |  *:              CW20        <L0001082>
   |    *:              W08         '電力
   |    *:              W20         '会社
   |*:              PRD         '.
  C: 0.000(0.08)  SE          <L0000020>
   |A: 0.002(0.09)  CW20        <L0004564>
   | |*:              CW08        <L0005172>
   | |  *:              P08         '小
   | |  *:              CW08        <L0000760>
   | |    *:              W22         '規模
   | |    *:              W08         '電力
   | |*:              W20         '会社
   |B: 0.998(0.91)  CW20        <L0004564>
   | |*:              CW08        <L0004122>
   | |  *:              CW22        <L0000214>
   | |    *:              P08         '小
   | |    *:              W22         '規模
   | |  *:              W08         '電力
   | |*:              W20         '会社
   |*:              PRD         '.
```

(b)

図11 解析木の例
Fig. 11 Examples of the obtained parsing trees.

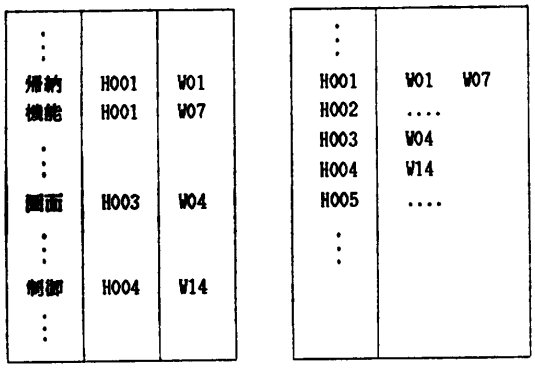
表 1 実験結果
Table 1 Results of the experimentation.

a.	訓練用データ数	28,568
b.	訓練用データの平均の長さ（単位：文字）	4.2
c.	テストデータ数	153
d.	正しい解析木に最大確率が付与されたテストデータの数	92
e.	正しい解析木に最大確率が付与されなかったテストデータの数	33
f.	パージングに曖昧さのなかったテストデータの数	22
g.	分割の誤りを含んでいたテストデータの数	6
h.	成功率 d/(d+e)	0.736

4. かな漢字変換エラーの自動検出に関する実験

本章では，漢字複合語の確率的構造解析の一つの応用として，かな漢字変換エラー検出の枠組みについて述べる．ここで，漢字複合語におけるかな漢字変換エラーとしては，ひらがな入力のミスや変換ミスにより，誤った漢字列が生成された場合を想定している．説明は，我々がそれに関して行った実験の，処理の手順に従って行う．なおこの実験の結果は，日本語文書校正支援システム CRITAC[13] の校正規則を実現する際に用いられている．

処理の手順は，次のとおりである．まず対象をJICST 電気工学編とし，そこから同音語を抽出して同音語のグループを作った．その際，一語から成るグループも許した．それらのグループには，H001, H002, … のような番号を付けておき，そのグループ分けに基づいて図 12（a）のような同音語辞書を作成し



(a) 同音語辞書 (b) 解析用辞書

図 12 同音語辞書と解析用辞書のフォーマット
Fig. 12 Formats of the two types of dictionaries.

た．同音語辞書の第1フィールドには短単位語の漢字のつづり，第2フィールドには第1フィールドの語が属する同音語グループの番号，そして第3フィールドにはその短単位語が属するクラスタの番号が記入されている．また，図 12 (b)に示すように，ひとつの同音語グループに属するすべての語のクラスタ番号が同時に引ける解析用辞書も用意しておき，構造解析の前処理で利用した．

以下では，かな漢字変換エラー「画面制御帰納」(本来なら「画面制御機能」) を例にとって，エラー検出の枠組みを述べる．まず，入力は，短単位語に分割されて，「画面・制御・帰納」の形で与えられる．これは，一般の漢字複合語解析の場合と同じである．次に，この入力に対して，同音語辞書による辞書引きが行われる．今の例の場合，辞書引き結果として，列「H003 H004 H001」が得られる．そのようにして得られた列が，構造解析の対象となる．まずこの列は，解析用辞書によって辞書引きされる．この例の場合，「きのう」に二つの漢字のつづりが存在するために，辞書引き結果として「W04 W14 W01」と「W04 W14 W07」の二つが得られる．これは，自然言語文の構文解析における，辞書引きの曖昧性とまったく同様のことである．

確率パーザは，列「H003 H004 H001」の構造として，計四通りの構造（上記の二通りの辞書引きおのおのに対して，それぞれ二通りの構造 ((H003 H004) H001) と (H003 (H004 H001)) が考えられる）のうち，どれが最も自然かを決定する．我々の手法は，正しくかな漢字変換された漢字複合語が，その成分である短単位語のクラスタ構成と，それらの間の係り受け関係，すなわち構造に関して最も高い確率を得るであろうという仮説に基づいている．

表 2 かな漢字変換エラー検出の実験結果
Table 2 Results of the experimentation of the automatic detection of Kana-to-Kanji misconversions.

a.	一語当り得られた解析木の平均個数	6.17
b.	最大確率の複合語が正しいつづりだったデータ数	54
c.	最小確率の複合語が誤ったつづりだったデータ数	70
d.	失敗（bでもcでもなかった場合）	22
e.	エラーを含んでいたデータ数	4
f.	成功率 c/(100−e)	0.73

ランダムに抽出した100語について，得られた結果を表2に示す．表2bは，入力が与えられたときに，同音語の可能性も考慮して得られるすべての解析木のうちで，最大確率を得たものが，正しい漢字列を導出していた場合の数を表している．ただし，ここでの同音語の可能性は，上でも述べたように，最初の入力漢字列の短単位語分割に対応した，ひらがな列の分かち書きに対してのみ考慮するものとする．表2cは，入力に対する可能なすべての解析木のうちで，最小確率を得たものが，誤った漢字列を導出していた場合の数を表している．

我々は，文中に現れた漢字複合語が，その読みに対する最小確率の解析木が導出する漢字列と一致したときに，その漢字複合語はかな漢字変換エラーを含むと指摘することを計画している．表2fの成功率は，このようにエラーを指摘したときに，そのメッセージが正しい（実際に変換エラーがあった）比率を表している．また表2dは，かな漢字変換エラーの指摘に利用できない解析結果の数を表している．

5. 現バージョンの問題点

漢字複合語の確率的構造解析における最大の課題は成功率の向上であるが，その問題の大部分は，文法の確率付けを改善することに帰着される．文法の確率付けを改善するためには，主に以下の二つの観点から検討を加える必要がある．

(1) 訓練用データを増やすこと
(2) 文法の修正

以下では，これらの項目の内容について説明する．

(1) エラー解析を行った結果，エラーのほとんどは，解析木の相対確率を，二文字語どうしの結合確率が決定してしまう場合であることが判明した．これは，複合語の結び付きに関する情報が少ないことを意

味する，したがって，成功率を向上させるためには，接頭，接尾辞を含む，複雑なより多くの漢字複合語データによる文法の確率付けが必要である．

（2） 現バージョンの文法では，複合語のカテゴリーは，その複合語を構成する最も右側の短単位語のカテゴリーによって決定されている．しかし，複合語の右端の接尾辞が，化，用，性，的，付，式などの接辞に近い語だった場合には，その前方の語のカテゴリーをその複合語全体のカテゴリーにした方が良いと考えられる．例えば複合語「技術的」のカテゴリーには，短単位語「技術」のカテゴリーを反映させるのが自然である．このことを実現するには，接辞に近い語の導出に関係する規則において，右辺第一項のカテゴリーを左辺のカテゴリーとするように変更すればよい．

6. むすび

漢字複合語の構造解析は，冒頭にも述べたように，より高度な日本語処理への鍵になると考えられる．その際，解析に必要な構文情報等を，人手で辞書に記述するのは，非常に困難である．そのような意味から我我は，この問題に対する確率的アプローチは，きわめて有効であると考えている．

漢字短単位語のクラスタリングにおいては，語の共起関係という構文的な情報だけを用いて得られたクラスタが，主に同意語，反意語，関連語等，意味的に関係がある語によって形成されていた．この事実は，複合語解析に対する確率的アプローチの有望な可能性を示している．今後，解析の成功率を向上させて行くことが，最大の課題である．

本稿で述べた漢字複合語解析の手法は，その前提となる複合語分割処理が，漢字複合語のみならず一般の複合語をも分割の対象としているならば，一般の複合語解析にも適用することができる．

謝辞 本論文をまとめるにあたり，貴重な御助言をくださった日本アイ・ビー・エム(株)東京基礎研究所，諸橋正幸氏ならびに，本研究を進めるにあたり多くの御助力をいただいた同，武田浩一氏，鈴木恵美子氏に深く感謝いたします．

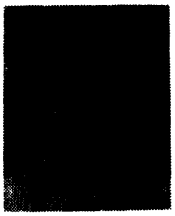
参考文献

1) 武田浩一，藤崎哲之助：統計的手法による漢字複合語の自動分割，情報処理学会論文誌，Vol. 28, No. 9, pp. 952-961 (1987).
2) 宮崎正弘：係り受け解析を用いた複合語の自動分割法，情報処理学会論文誌，Vol. 25, No. 6, pp. 970-979 (1984).
3) 西野哲朗：漢字複合語の確率的構造解析のための漢字短単位語基のクラスタリング，第32回情報処理学会全国大会論文集，pp. 1569-1570 (1986).
4) 西野哲朗，藤崎哲之助：漢字複合語の確率的構造解析の試み，第 34 回情報処理学会全国大会論文集，pp. 1193-1194 (1987).
5) 野村雅昭：複次結合語の構造，国立国語研究所報告 49，電子計算機による国語研究 V，pp. 72-93 (1973).
6) 野村雅昭：三字漢語の構造，国立国語研究所報告 51，電子計算機による国語研究VI，pp. 37-62 (1974).
7) 野村雅昭：四字漢語の構造，国立国語研究所報告 54，電子計算機による国語研究 VII，pp. 36-80 (1975).
8) 石井正彦，野村雅昭：機械工学用語の語種構造，計量国語学，Vol. 14, No. 4, pp. 163-175 (1984).
9) 藤崎哲之助：自然言語の曖昧さの取り扱いの研究，東京大学大学院情報工学科博士論文 (1985).
10) Fujisaki, T.: An Approach to Stochastic Parsing, *Proc. of COLING 84*, pp. 16-19 (1984).
11) 奥野忠一ほか：続 多変量解析法，p. 299，日科技連，東京 (1976).
12) Baum, L. E.: An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes, *Inequalities*, Vol. III (1972).
13) 鈴木恵美子，武田浩一，藤崎哲之助：日本語文書校正支援システム CRITAC，情報処理学会日本語文書処理研究会資料，8-5 (1986).

(昭和 63 年 5 月 19 日受付)
(昭和 63 年 9 月 5 日採録)

西野 哲朗（正会員）

昭和 34 年生．昭和 57 年早稲田大学理工学部数学科卒業．昭和 59 年同大学大学院理工学研究科博士前期課程修了．同年日本アイ・ビー・エム(株)入社．現在，東京電機大学理工学部情報科学科助手．この間，属性文法，確率的自然言語解析の研究に従事．計算量理論，言語理論，およびオートマトン理論に興味を持っている．電子情報通信学会，日本ソフトウェア科学会，CAI 学会各会員．



藤崎哲之助（正会員）

昭和 22 年生．昭和 44 年東京大学工学部計数工学科卒業．昭和 46 年同大学院修士課程修了．同年日本アイ・ビー・エム(株)に入社．同社東京基礎研究所にて自然言語処理，人工知能の研究に従事．現在 IBM トーマス・J・ワトソン研究所にてパターン認識の研究に従事．工学博士．電子情報通信学会，日本ソフトウェア科学会，計量国語学会，ACM 各会員．