論　文

帰納的学習を用いたタグなしコーパスからの統語規則の自動獲得手法

渋木　英潔[†]　　荒木　健治[†]　　栃内　香次[†]

Automatic Acquisition Method of Syntactic Rules from Untagged Corpora Using Inductive Learning

Hideyuki SHIBUKI[†], Kenji ARAKI[†], and Koji TOCHINAI[†]

あらまし　本論文では，帰納的学習を用いて，タグなしコーパスから統語上の解析に必要な規則（統語規則）を自動的に獲得する手法について述べる．統語規則を事前に人手で与えることはばく大な労力が必要である．また，人手によって作成された統語規則は静的（固定的）なものになり，現実の文のような様々な言語現象に対処することができない．そこで，多くの文を処理できる統語規則は正しいという制約のもと，類推と統計的基準に基づいた帰納的学習を用いて統語規則を動的に獲得することにより，対象に動的に適応する手法を開発した．本手法は，以下の三つの点で従来手法よりも頑健である．第1に，分かち書きも，いかなるタグも付けられていない文から統語規則を獲得できる．第2に，語彙，品詞，文法はすべて空の状態から開始できる．事前に備えているのは，文字という言語単位の存在と，統語規則などの表現形式と，統語規則の運用・学習方法だけである．第3に，単一の機構で，人間の言語獲得過程におけるすべての時期の文（1語文から3語文以上まで）を処理できる．本手法を実装したシステムを作成し，統語規則のない状態から，外国人のための日本語学習用テキストを30回繰り返して入力した結果，獲得された統語規則を用いて85.3%の解析成功率が得られた．また，解析が成功した結果の42.8%が正解であることを確認した．

キーワード　自然言語処理，帰納的学習，文脈自由文法，形態素解析，構文解析

1. まえがき

自然言語を計算機で処理する場合の一手法として，規則に基づいて解析する方法がある[1],[2]．この解析に使用する規則（以降，統語規則と呼ぶ）が詳細に記述されているほど，精度の高い解析が行われることが期待できる．しかしながら，統語規則を事前に人手で与えることはばく大な労力が必要となり，また，人手で作成された統語規則は静的（固定的）なものとなり，現実の文のような様々な言語現象に対処することができない等の問題点がある．この問題を解決するために，本論文では，帰納的学習を用いて統語規則を自動的に獲得し，対象に動的に適応する手法を提案する．本論文における帰納的学習とは，類推と統計的基準に基づいて，ある実例を解析するために不足している規則を補う過程のことである．

本論文において，獲得される統語規則は，語彙と文脈自由文法の規則である．獲得される統語規則は，本質的には文脈自由文法である必要はないが，文構造を表現する上で文脈自由文法が基礎として広く一般に使われていることから，文脈自由文法を対象とした．本論文では，この統語規則を蓄積したものを統語規則辞書と定義する．また，本論文では日本語を対象としているが，本手法のアルゴリズムは日本語特有の言語現象に依存するものではない．

以下，本論文では，研究の背景，基本的な考え方を述べた後，本手法の処理過程と本手法を実装したシステムの評価実験について述べる．

2. 背　景

自然言語は人間が社会的な生活を営む上で半ば必然的に生まれたものである[2]以上，人間の言語処理を模倣することは自然言語を処理する上でも有効であると考えられる．人間の言語処理における大きな特徴として学習が挙げられる．我々は，これまでに帰納的学習を用いて未知語の獲得を扱った研究を行い，帰納的

[†] 北海道大学大学院工学研究科電子情報工学専攻，札幌市
Division of Electronics and Information Engineering, Graduate School of Engineering, Hokkaido University, Kita 13, Nishi 8, Kita-ku, Sapporo-shi, 060-8628 Japan

学習が語彙の獲得に有効であることを示した[3]．本論文は，この研究を発展させ，帰納的学習を用いた統語規則の獲得に適用する．以下で，統語規則を獲得する従来研究と比較しながら，本手法を設計する上での基本思想を述べる．

統語規則を獲得する従来研究には文献[4]～[10]などがある．文献[4]～[6]は，品詞タグや構文情報が付与されたコーパスを使用し，そこから得られる情報を最大限活用して統語規則を獲得する．コーパスに基づいて統語規則を獲得する場合，学習材料となる文の量が多いほど，様々な表現の文が入力されることになり，現実の文への適応性が高い統語規則を獲得できる．そのため，現実の言語表現をすべて網羅するような普遍的なコーパスが与えられることが望ましい．しかしながら，地域，社会，時代，使用者によって異なる自然言語[11]において，普遍的なコーパスを作成することは極めて困難であると考えられる．現実問題として，状況に応じて複数のコーパスを与えていくことが必要となる．しかしながら，コーパスごとに品詞体系が異なることや，それに伴う単語境界の認定基準も異なることが多く[12]，字面以外の情報を共有して用いることは困難であると考えられる．以上の理由から，頑健なシステムには，分かち書きされておらず，品詞及び構文情報などの字面以外の情報をもたない文を入力とできることが必要である．

タグなしコーパスから統語規則を獲得する研究として，文献[7]～[10]が挙げられる．文献[7]は，事前に与えられた品詞集合から，作成可能な統語規則を作成する．その後，Inside-Outsideアルゴリズムを用いて，確率的にふさわしい統語規則を選択する．文献[4]～[7]に共通した特徴として，解析を行う前に学習用コーパスから統語規則を獲得する点が挙げられる．解析を行っている間は，解析結果がどうであれ，統語規則に影響を与えない．つまり，獲得される統語規則は，静的なものとなり，学習用コーパスの特徴に大きく左右される．

上述のように，普遍的な学習用コーパスの作成は困難であり，したがって，入力環境の変化に応じて動的に適応できることが必要である．

そのような動的に適応する問題を扱った研究として，文献[8],[9]がある．文献[8],[9]では，基本的な統語規則の集合を事前に与え，その統語規則を用いて解析した結果から新たな統語規則を獲得する．すなわち，既に存在する統語規則を新たな領域に適応させることを目的としており，本手法のようにすべて空の状態から獲得することを対象としていない．文献[8]の実験結果に見られるように，彼らの手法では，獲得するために十分な規則が事前に与えられなければならない．獲得の性能が，事前に与えられた規則に大きく影響を受けることは，頑健なシステムとしてふさわしくない．頑健なシステムには，統語規則の初期状態が空の状態からでも統語規則の獲得が可能であることがふさわしい．本論文において，統語規則の初期状態を空とするということは，単純に統語規則が存在しないということだけでなく，統語規則を構成するために必要な語彙や品詞も与えないということである．例えば「出発する」という言葉の品詞を考える．これを「出発する」という1語の動詞と見るか，「出発」という名詞に「する」というサ変動詞が付いたものと見るかは，人によって異なる．このようなあいまい性が存在することから，我々は，品詞を含む統語的な知識は，人間が生得的に備えているものではなく，現実の文の中から独自に獲得するものであるという立場をとっている．そのため，統語規則の存在は仮定するが，品詞をはじめとして具体的な統語規則を事前に与えることはしない．ただし，本手法は，計算機上でテキストを入力して処理するため，最小の言語単位が1文字であることは事前に知っているものとする．

本手法と同じように，品詞を含む統語規則を学習の対象とした研究として文献[10]がある．文献[10]では，言語の獲得過程を学習一般に共通する枠組みの中で説明することを目的として，言語と外界の2種類の情報を処理し，それらの対応関係をもとに統語規則と意味概念を同時に獲得することを行っている．しかしながら，対象としているのは，言語獲得の初期（1語文から2語文獲得の時期）段階であり，言語獲得の後期に見られる比較的複雑な統語構造の獲得を扱っていない．そのため，獲得される統語規則は単純なものである．処理できる文の種類が限られていることは，頑健なシステムにふさわしくない．したがって，本手法では，処理の対象となる文の種類を限定しない．

以上から，本手法は，以下の三つの点において従来手法よりも頑健である．

- 分かち書きも，いかなるタグも付けられていない文から統語規則を獲得できる．
- 語彙，品詞，文法はすべて空の状態から開始できる．
- 処理対象となる文の種類を限定しない．

本手法は，以上の点をすべて満たし，かつ，入力に動的に適応することを目的としている．

3. 基本的な考え方

2. を背景として，基本的な考え方について述べる．以下では，本手法の全体の流れを述べた後，統語規則と解析結果の表現形式，そして，帰納的学習のアルゴリズムを述べる．

3.1 全体の流れ

全体の流れを図 1 に示す．まず，分かち書きされず，タグも付けられていない入力文に対し，今までに獲得された統語規則を用いて，形態素解析と構文解析を行う．本手法の解析の目的は，入力文が非文かどうかを判断するものではなく，今までに獲得された統語規則を用いて，帰納的学習に必要な品詞や係り受け関係などの情報を付与することである．解析が完全に成功しない場合には，解析に必要な統語規則が欠如していると判断する．また，入力文には，意味的な非文（内容を伝えることのできない文）はないと仮定する．そのため，どのような入力文に対しても必ず何らかの解析結果を導き出すことができる．一般に，解析結果は複数導き出される．システムは，事前に与えたゆう度評価の式に従って，それぞれの解析結果に対してゆう度評価を行う．ゆう度評価が 1 位以外の解析結果はすべて破棄し，唯一の解析結果だけを残す．これにより，

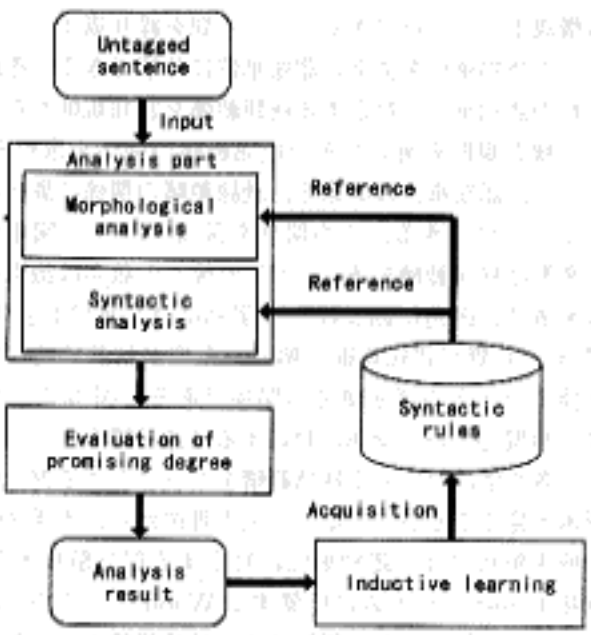

図 1 処理過程
Fig. 1 Outline of process.

獲得される統語規則を適切なものに収束させることができる．人間の場合にも，多義語の意味を理解する過程において瞬間的に複数の結果が浮かんだ後，唯一の結果に絞られることが知られている [13], [14]．我々は，これと同様のことが統語上の解析においても起こっていると推測する．

解析結果は解析木で与えられる．ゆう度評価 1 位の解析結果が不完全である場合，解析に必要な統語規則が欠如していると判断する．不完全な解析結果とは，解析木の頂点のノードの数が 1 でない場合か，解析木の品詞に相当するノードと語彙とが結び付いていない場合のどちらか，または，両方の場合である．不完全な解析結果から，事前に備えている帰納的学習のアルゴリズムに従って，不完全な箇所を補うような統語規則を獲得する．今回の入力文から得られた解析結果や統語規則と，過去の入力文から得られた解析結果や統語規則を比較して，統語規則全体の最適化を図る．その後，次の入力文に移る．

3.2 統語規則の表現形式

本手法において，文脈自由文法はチョムスキー標準形 [1] で獲得される．すなわち，A, B, C を非終端記号，α を終端記号としたとき，獲得される文脈自由文法は下の二つの式で表される．

$$A \rightarrow B\ C \tag{1}$$

$$A \rightarrow \alpha \tag{2}$$

本手法において，(1) の規則を統合規則（unitive rule）と呼び，(2) の規則を認知規則（perceptive rule）と呼ぶ．A, B, C を統語範疇（syntactic category）と呼び，識別番号で表現する．日本語処理では，一般に α に形態素が入る．しかしながら，日本語の認知においては，文節よりも長い単位が一度に検出されることがあり，必ずしも形態素単位で認知しているわけではないことが報告されている [15]．文献 [15] では，この単位を認知単位と呼び，認知単位を基本とした解析法は，形態素を基本とした解析法よりも効率が良いことを示している．我々は，認知単位が，人間が日本語を効率良く処理する上で独自に形成した単位であると考える．ゆえに，本手法では，α の単位を形態素と規定せず，処理効率の良い自由な単位で獲得することにした．α の単位を認知単位（perceptive unit）と呼ぶ．

3.3 解析結果の表現形式

本手法は，3.1 で述べたように，入力文の解析結果をもとに帰納的学習を行う．不完全な箇所を補うよう

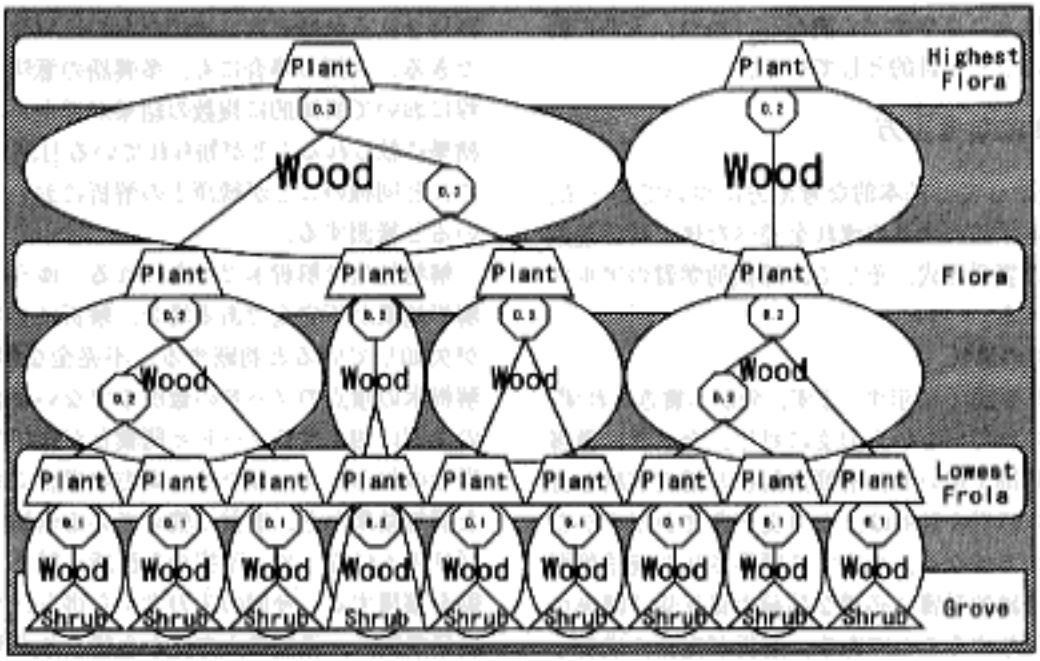

図 2　Forest の概略図
Fig. 2　Image of Forest.

に学習することから，既に獲得された統語規則では，入力文のどの部分が解析できないのかを明らかにできる解析結果でなくてはならない．更に，可能ならば，どのような統語規則があれば解析できるのかを推測できる解析結果であることが望ましい．規則に基づいて解析する場合，解析結果は一般に一つの解析木で表現される．しかしながら，上に述べた条件を満たすためには，一つの解析木による表現では不可能である．そのため，本手法の解析結果は，小さな部分解析木の集合として表現される．この部分解析木を Wood と呼び，Wood の集合で表される全体の解析結果を Forest と呼ぶ．Forest の概略図を図 2 に示す．図全体が Forest であり，円で囲まれた部分解析木それぞれが Wood である．

自然言語処理では一般に，入力は，形態素解析，構文解析，意味解析の順に解析され，各解析ごとに解析結果が求められる．各解析の間に存在する解析結果は，次の解析に受け渡すための中間生成物であり，最終的な解析結果の一要素であることが多い．構文解析へつなぐ一般的な日本語形態素解析では，入力文の単語境界を明らかにし，対応する品詞を求めている（例えば，[16]）．しかしながら，これら形態素解析の結果に含まれる情報は，解析木の中で終端記号と前終端記号のペアにより表現できる．前終端記号とは，終端記号の一歩手前の記号であるという意味で，文脈自由文法における品詞の別称である [2]．また，本手法では形態素の代わりに認知単位を用いることを述べた．例えば，「無関心だが」という認知単位が獲得されていると仮定する．この認知単位は，「無関心」と「だが」という二つの認知単位から構成することが可能である．更に，「無関心」は「無」と「関心」の二つの認知単位から構成することができる．この手順を繰り返すと，すべての認知単位を文字の認知単位に分解できる．それぞれの認知単位に対応する統語範疇を認知規則から求め，統合規則を用いてその統語範疇の関係を求める．すると，認知単位の関係を，統語範疇の関係に置き換えることができる．この関係を図 3 に示す．図中の八角形は統語範疇を表し，"T" の後ろの数字は識別番号を表す．四角に囲まれた文字列は認知単位を表す．図 3 の右側が認知規則の例で，左側が統語範疇に置き換えたときの認知単位の関係である．図 3 の左側は，「無関心だが」全体に対応する統語範疇をルートとし，各文字に対応する統語範疇をリーフとする部分解析木と見ることができる．以上の理由から，本手法の形態素解析では，認知単位に対応する部分解析木を，直接 Forest のボトムに位置する Wood として当てはめていく．認知規則は認知単位と統語範疇の他に部分解析木の情報を保持する．この部分解析木を Shrub と

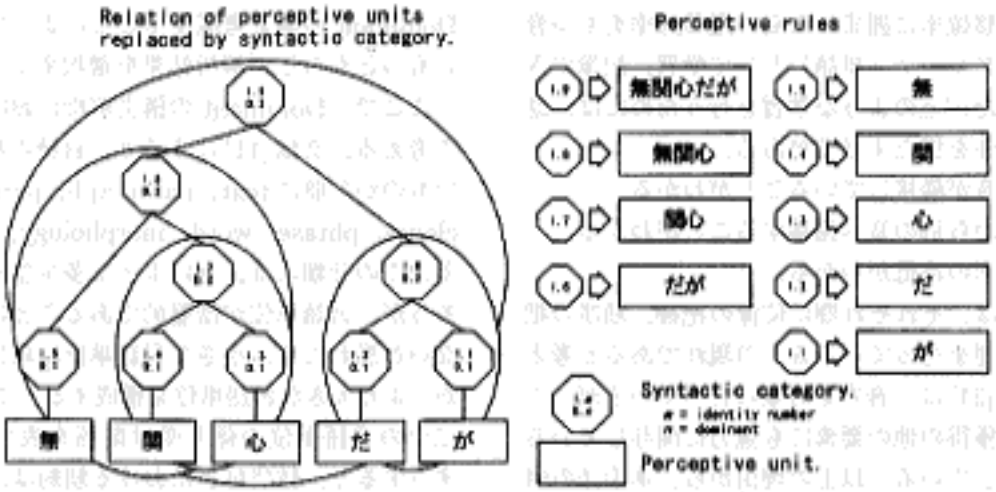

図 3　認知規則と統語範疇で置き換えた認知単位の関係
Fig. 3　Perceptive rules and the relation of perceptive units replaced by syntactic category.

呼ぶ．図 3 の左側においては，円に囲まれた部分それぞれが Shrub を表している．認知規則と Shrub を結ぶ矢印は，対応関係を表しており，認知規則は，矢印で結ばれた Shrub を保持する．図中では省略しているが，1 文字の認知単位をもつ認知規則も自分自身だけからなる Shrub を保持している．また，形態素解析を行った時点で作成される Shrub の列で表される Forest の一部を Grove と呼ぶ（図 2 参照）．三角に重なる Wood が Shrub を表している．底辺に位置するそれらの Shrub の列が Grove である．このように，形態素解析の段階で，Forest の一部を作成することにより，効率良く構文解析へとつなげることができる．

3.4　帰納的学習

本手法は，統語規則辞書がすべて空の状態から，帰納的学習を用いて，入力文から帰納的に統語規則を獲得する．しかしながら，人間が獲得する知識の中には，外的入力だけからは帰納できない部分が存在することが示されている [17]．この外的入力だけから帰納できない部分は，人間が生得的にもっている知識によって，解決されていると考えられている．言語獲得の領域における生得的な知識はどのようなものであるかという問題に対して，現在，様々な意見が出されている．一つには，人間は生得的に統語規則の雛型である普遍文法をもっているという考え方がある [18]．この考え方によれば，言語獲得とは，周囲の言語入力から，普遍文法中のパラメータを調節する過程のことである．言語を獲得できるかどうかは，普遍文法をもっているかどうかによる．一方，言語訓練を受けさせることで霊長類がある種の「言語」を使用できるようになった例 [11], [19] が確認されている．もちろん，それらの言語は人間の言語と同一のものではない．文献 [19] は，この違いがメタ言語のような論理的階層性を用いることができるかどうかによるものだと主張する．また，人間も動物の一種であり，他の種と同様に，人間の機能も進化の過程で獲得されたものであるという基本的な生物学的認識を忘れてはならないとも述べている．この考え方では，言語が獲得できるかどうかは，学習機構の違いによる．我々は，後者の立場に立ち，言語のどのような側面に着目し，どのような基準で判断するかといった学習方略に関する知識だけをもつという立場をとった．

本手法の帰納的学習は，多くの文を処理できる規則は正しいという制約を根底にもっている．その制約を実現するために，位置と順序関係に着目して類推を行い，類推によって解決できない場合は統計的基準で判断する．類推が，言語活動において極めて重要な役割を果たしていることは一般に知られており，類推を自然言語処理システムに応用した研究も数多い（例えば，[20]）．また，位置と順序の関係に着目して統計的基準を用いる根拠としては文献 [21] がある．文献 [21] は，8 か月の幼児が，隣接する音の間の関係を統計的に処理する能力を備えていると主張し，この能力によって得た 2 音間の遷移確率を使用して，連続音声から単語分割を行っていることを実験により確認した．また，

幼児が低い遷移確率に囲まれた高い遷移確率をもつ音の列を特別なグループ（単語）として学習，記憶できることを示した．上のような学習を行うためには，幼児が以下の条件を満たす必要がある．

- 二つの音が隣接していることがわかる．
- ある音から別の音へ遷移することがわかる．
- 遷移確率の高低がわかる．

これらの条件は，それぞれ順に位置の把握，順序の把握，統計的処理を行っていることの現れであると考えられる．文献[21]は，音声を対象に行っているが，この能力が言語獲得の他の要素にも強力に関与している可能性も示唆している．以上の理由から，本手法の帰納的学習は，位置と順序関係に着目し，類推と統計を判断基準とする．本手法の帰納的学習のアルゴリズムを以下に示す．

[帰納的学習のアルゴリズム]

（1） 周囲の環境が同一である統語範疇は同一である．

（2） 隣接する統語範疇は一つの統語範疇にまとまる．

（3） 現実の文に多く出現する統語規則は正しい．

（4） 同じ位置関係にある統語範疇と認知単位は結び付く．

（5） 隣接する認知単位は一つの認知単位である可能性がある．

以上のアルゴリズムの具体的な用法は4.の中で，状況に応じて述べる．

4. 処理過程

4.1 形態素解析

本手法の形態素解析は，多段階分割法[22]に基づいたものである．多段階分割法は，高い確実性をもつ単語を用いて一意に決定できる部分から決定し，その後も，確実性の高い部分から順次決定していく手法である．この手法は解析のあいまい性を減少させるのに有効である．文献[22]では確実性の高いキーワードを用意しているが，本手法では，すべての統語規則を帰納的学習により獲得するため，事前に用意しておくことはできない．獲得される統語規則の確実性を表すためにDominantの概念を導入する．Dominantは，最も低い確実性を1とした数値で表される属性である．すべての統語範疇はDominantをもつ．また，文字に対応する統語範疇のDominantは1である．形態素解析において複数の解析結果が求められる場合に，Dominantの値を考慮することによって，確実性の高いもっともらしい解析結果を選択することができる．

ここで，Dominantの構文解析における影響について考える．文献[11]によると，言語の単位には，大きいものから順にtext, paragraph, period, sentence, clause, phrase, word, morphology, phonemeがある．この分類には，人によって多少差異が存在するだろうが，言語単位が階層的であることについて異論はないと思われる．小さな言語単位は結び付いて，同じか，より大きな言語単位を構成する．この結び付きを，二つの言語単位の係り受け関係を表しているとみなす．すると，結び付きにおける制約は，構文解析の制約と考えることができる．そこで，結び付くことができる言語単位の条件を内省により求めると，基本的に同じ大きさの言語単位同士は直接結び付くことが判明した．異なる言語単位同士について考慮すると，例えば，morphologyとwordは直接結び付くことはあるだろうが，morphologyとphrase，あるいは，clauseが直接結び付くことは，まず考えられない．このことから，直接結び付くことができる言語単位には，両者の大きさが，大きくかけ離れていないことという制限があると推測できる．本手法では，Dominantを言語単位の大きさと一致させて処理を行う．したがって，統合規則 $A \rightarrow BC$ を構成する統語範疇のDominantを以下のように定義する．A, B, C のDominantを，それぞれ a, b, c としたとき，a, b, c には次の関係が成り立つ．

[Dominantの定義]

（1） b と c の差は1以内である．

（2） b と c が同じ場合，a は b, c よりも1高い．

（3） b が c より高い場合，a は b と同じである．

（4） c が b より高い場合，a は c と同じである．

図2と，図3の左側において，統語範疇に書かれた"D"の後ろの数字が，その統語範疇のDominantを表す．図中のDominantが上の定義による関係を満たしていることに注意してもらいたい．

入力文中の，まだ，Shrubを当てはめられていない部分を未定範囲と呼ぶ．形態素解析では，未定範囲に対して認知規則を，その統語範疇のDominantの高い順に，適用できるか調べていく．認知規則の認知単位が未定範囲の中に含まれている場合，認知規則を適用できる．ただし，Dominantの定義（1）から，その適用しようとする範囲の直前と直後に既にShrubが当てはめられており，かつ，そのShrubのルートの

Dominantがともに，適用しようとする認知規則の統語範疇のDominantより2以上高い場合には適用することができない．Dominantの定義（1）は2以上のDominant差がある統語範疇を統合する規則が存在しないことを示す．そのため，仮に，そのようなShrubを当てはめたとしても，構文解析において，それ以上統合される可能性がないからである．また，同じ未定範囲に対して適用できる同じDominantをもつ複数の認知規則が存在する場合，最も大きい範囲に適用できる認知規則を選択する．解析結果は4.3に後述する式(3)に従ってゆう度評価され，高いゆう度をもつ解析結果だけが残される．ゆう度は，Dominantと適用される範囲の文字数に影響を受ける．同じDominantであれば，適用される範囲の文字数の多い方が高いゆう度となる．したがって，Dominantによる基準で解決できない場合は，適用される範囲の文字数を基準にして解決する．しかしながら，認知単位の適用範囲が交差している場合に限り，交差しているすべての認知単位を適用して解析結果を求め，最終的な解析結果の判断をゆう度評価に委ねる．これは，「ここではきものをぬいでください」における「ここでは」と「はきもの」のような例が考えられるからである．以上の手順を繰り返し，最後まで未定範囲が残った場合，未定範囲ごとに一つの未知語とみなす．入力文に対して，一つもShrubを当てはめることができなかった場合，入力全体が未定範囲となり一つの未知語とみなされる．したがって，統語規則辞書が空の状態では，適用できる認知規則が存在しないため，入力文を一つの未知語としてしか認知できない．その結果，形態素解析では，入力文を分割することなく，一つの未知語だけからなる解析結果を唯一の解析結果として，次の構文解析に送る．

4.2 構文解析

本手法は，事前に具体的な統語規則を与えず，入力文が必ずしも統語的な文単位で行われるとは限らず，未知語の存在を許した中で，統語規則を獲得することを目的としている．ゆえに，ルートやリーフとなる統語範疇に完全依存した解析手法は，その性能を完全に発揮することができない．このような従来の基本的手法の問題点を指摘した構文解析の研究として文献[23]がある．しかしながら，文献[23]は，この問題を解決する手段として，意味情報や機能語を利用して補う方法をとり，それらの知識は事前に備えているものとした．そのため，解析に必要な規則を獲得する（つまり，獲得するまでは解析に必要な規則が存在しない）状況で動作することを想定して設計されたものではない．この問題を解決するために，我々はWood単位で部分解析木を作成し，Woodを接合していくことでForestを構成するGraft法[24]を開発した．以下に，Graft法とGraft法を用いた問題解決の概略を述べる．

未知語の問題は，Woodを作成するときに解決する．Woodのルートを Trunkと呼び，リーフを Twigと呼ぶ．Twigに依存しないために，Wood内部はトップダウンに展開する．トップダウン法の出発点となるTrunkを明らかにするために，Reachable List（RL）の概念を導入する．RLはDominantと同様にすべての統語範疇ごとに存在し，各々の統語範疇から到達可能な統語範疇をもとに構成される．到達可能かどうかは次のように判断する．

- 統語範疇が自分自身に到達可能である．
- 統合範疇 $A \rightarrow B\ C$ から，B は A のRLに含まれるすべての統語範疇に到達可能である．

ゆえに，最も左に位置するTwigはTrunkに到達可能である．この最左Twigを，特別にGraftと呼ぶ．ある統語範疇のRLには，統語範疇からDominant差が1以内である到達可能な統語範疇だけが含まれる．したがって，Woodは次の手順で作成される．

- GraftのRLを用いてTrunkを予測する．
- Trunkから，Graft以外のTwigをトップダウンに導き出す．

統語的な文単位で入力が行われない，つまり，Forestのルートが不明である問題は，Woodを接合するときに解決する．各々のWoodは横型ボトムアップ法で接合する．Woodを接合する段階では，ルートとなる統語範疇に依存せずに，現在の統語規則で可能な限りの処理を行う．このとき，接合点となっている統語範疇をPlantと呼び，Plantの列をFloraと呼ぶ（図2参照）．最上層Floraに含まれるPlantの数が1であれば，入力全体に対する解析木が作成できたことになる．以上のようにして，入力文のどの部分が解析できないのかを明らかにでき，かつ，どのような統語規則があれば解析できるのかを推測できる解析結果を作成する．

統語規則辞書が空の状態では，構文解析に渡された形態素解析の結果は，一つの未知語で表現されている．したがって，これ以上，統合されることなくゆう度評価に渡される．

4.3 ゆう度評価

ゆう度評価は，形態素解析結果Groveと構文解析結

果 Forest に対してそれぞれ行われる．

$$PDG = \frac{1}{NS}\sum_{i=1}^{NS}(DS_i \times NCS_i) \quad (3)$$

PDG（Promising Degree of Grove）は Grove のゆう度である．NS（Number of Shrub）は Grove 中の Shrub の数である．DS_i（Dominant of Shrub）と NCS_i（Number of Character in Shrub）は，それぞれ，i 番目の Shrub のルートである統語範疇の Dominant と，リーフである認知単位の文字数である．この式は，入力文をできるだけ少ない Shrub に分割し，各 Shrub ができるだけ多くの文字を高い Dominant をもつ統語範疇でまとめている場合に，高い PDG を導き出す．PDG は，形態素解析が行われた時点で求められる．システムは PDG に従って，上位 10 位までの Grove を構文解析に送る．10 位よりも下の Grove は，Dominant の低い Shrub で構成されていることが多い．そのため，構文解析に送っても，高いゆう度評価を得られる Forest を作成できないと判断した．

作成された Forest は，最上層 Flora の Plant に基づいてゆう度評価が行われる．

$$PDF = \frac{1}{NP}\sum_{i=1}^{NP}(DP_i \times NCP_i) \quad (4)$$

PDF（Promising Degree of Forest）は Forest のゆう度である．NP（Number of Plant）は最上層 Flora 中の Plant の数である．DP_i（Dominant of Plant）と NCP_i（Number of Character corresponding to Plant）は，それぞれ，i 番目の Plant の Dominant と，その Plant が統合している文字数である．この式は，できるだけ少ない Plant に統合し，各 Plant ができるだけ多くの文字を高い Dominant をもつ統語範疇でまとめている場合に，高い PDF を導き出す．PDF と PDG の和が，その Forest の最終的なゆう度評価となる．最終的なゆう度評価が最も高かった Forest が，入力文に対する解析結果である．

4.4 統語範疇の獲得

統語範疇の獲得に関する従来研究には，統語範疇自身を獲得するのではなく，人手により事前に与えられた統語範疇に語彙を分類するものがある（例えば，[25]）．しかしながら，文献 [25] のように，その統語範疇の分類を議論せずに決定することが多く，分類の明確な根拠があるわけではない．

我々は，2. で述べたように，統語範疇の体系を入力文から自動的に構築していく立場をとっている．本手法では，すべての統語範疇を識別番号で表現しているため，必要に応じて任意に獲得することができる．新規に獲得する統語範疇には，新規の識別番号が割り当てられる．獲得する統語範疇の種類が無制限に増えることを防ぐため，3.4 の帰納的学習のアルゴリズム（1）に従って，複数の統語範疇を一つの統語範疇に同一化する．例えば，「太郎/は/札幌/に/行った」と「太郎/は/小樽/に/行った」という二つの文を比較する．"/" は単語境界を示す．この例では，「札幌」と「小樽」以外は種類，出現位置ともにすべて同じである．この場合「札幌」と「小樽」は同じ統語範疇に属すると推測する．便宜上，字面レベルで説明したが，実際には統語範疇に置き換えて比較する．しかしながら，現実の文において，このように文全体がマッチングする可能性は低いと考えられる．そこで，Forest が Wood の集合であることを利用して，Wood 単位で比較を行う．先の例でいうと，これは「札幌/に」と「小樽/に」の連用句単位で比較することに相当する．Wood 単位の比較は，文全体の比較よりもマッチングの確率が上がる．ただし，比較できるのは同一の Trunk をもつ二つの Wood とした．これは，文と文，連用句と連用句の比較は行えるが，文と連用句の比較は行えないことを意味する．以上のようにして，自動的に品詞を含む統語範疇の体系を構築していく．

4.5 統合規則の獲得

最上層 Flora に存在する Plant の数が 1 でない場合，その Forest は不完全である．この場合，3.4 の帰納的学習のアルゴリズム（2）に従って，それらの Plant で示される統語範疇を統合する統合規則を獲得する．また，4.1 の Dominant の定義（1）から，2 以上 Dominant 差のない隣接する Plant であることが追加条件となる．しかしながら，Plant の数が 3 以上存在する場合，中央の統語範疇が左右のどちらの統語範疇にかかるのか，一意に決定することはできない．また，一度条件を満たしたからといって，即座にその統語範疇を統合する規則を獲得することは妥当ではない．この問題は，3.4 の帰納的学習のアルゴリズム（3）に従って解決する．例えば，…XAY…という統語範疇の並びがあるとする．A に対して XA と AY の二つの統語範疇のペアをとり，過去の入力文から，どちらのペアが多く出現しているか統計をとる．一方のペアがしきい値以上出現した場合，そのペアを

右辺にとり，左辺の統語範疇を新規に獲得して，統合規則を獲得する．このとき，別の問題が起こる．本手法が，無限の入力を仮定しているため，単純に出現頻度を数えた場合，無限の入力の中では，間違いのペアであってもいずれはしきい値を超えてしまう．この問題を解決するために活性値の概念を導入する．活性値は，統合規則ごとに備えられており，数値で表される．まず，新規のペアが出現した時点で，しきい値未満の活性値をもつ統合規則を獲得する．しきい値未満の活性値しかもたない統合規則は，解析において使用することはできない．その後，同じ並びのペアが出現するたびに，1よりも大きい値だけ活性化させる．この値は前後の文脈や獲得される統合規則の重大性などにより変動することが考えられるが，今回は最初の研究であることから固定値とした．また，しきい値を求めるために予備実験を行った．予備実験の結果から，しきい値を高く設定するほど，正しい解析結果が得られる統語規則を獲得することができるようになるが，その代わりに，学習速度が遅くなることが判明した．したがって，最終的なしきい値は，解析精度と学習速度のバランスをもとに，ユーザの意図に委ねられると考えられる．今回の論文では，予備実験の結果をもとに，解析精度と学習速度をともにある程度満たす値として，しきい値を新規のペアに与えられる活性値の6倍に設定した．しきい値以上に活性化された統合規則は，解析で使用可能になる．活性値は，活性化されたかどうか，統合規則が使用可能かどうかにかかわらず，入力が行われるたびに1ずつ減少していく．その結果，活性値がしきい値未満に減少した統合規則は解析で使用できなくなる．また，活性値が0になった統合規則は統語規則辞書から削除される．このことは，人間が長い間使用しなかった知識を忘却していくことに類似している．人間の忘却は，時間によって記憶痕跡が減衰していくとも，精神活動によって記憶に干渉が生じることによるものだとも考えられている[26]．しかしながら，その正確な原因は特定されていない．本手法のような自然言語処理システムにおいては，厳密に時間に依存して忘却していくことは処理上の有効性がない．そのため，入力文を忘却の単位とした．

4.6 認知規則の獲得

統語規則と結び付いていない部分文字列が存在する場合，そのForestは不完全である．構文解析によって，その部分文字列の品詞に相当する統語範疇が推測されている場合，3.4の帰納的学習のアルゴリズム（4）に従って，その部分文字列を認知単位とし，品詞に相当する統語範疇と結び付ける認知規則を獲得する．品詞に相当する統語範疇が不明の場合，その部分文字列を文字に分解する．その中にいまだ獲得していない文字が存在する場合，新規に統語範疇を獲得し，その文字と結び付ける認知規則を獲得する．したがって，統語規則の少ない初期の段階では，文字を認知単位とした認知規則しか獲得できない．以下に2文字以上の認知単位を獲得していく過程を説明する．2文字以上の認知単位は，3.4の帰納的学習のアルゴリズム（5）に従って獲得する．「可能性」問題については，統合規則と同様に活性値を認知規則に導入することで解決する．Forestから，二つのShrubのルートを統合している統合規則を探す．その統合規則の左辺の統語範疇と，両方のShrubに対応する認知単位を複合した認知単位を結び付ける認知単位を獲得する．複合認知単位が獲得されるのは，Shrubを結び付ける統語規則が獲得された後である．また，認知単位に含まれる文字数が多くなるほど，出現頻度が低下することから，文節を超える長さをもつ認知単位が獲得されることはまれであると考えられる．

5. 評価実験

本手法を実装したシステムを作成し，評価実験を行った．表1に示す文献から，タグなしコーパスを作成し入力した．また，同様の領域に属すると判断した文献ごとにグループ分けをした．今回の実験では，人間の成長過程を考慮に入れて，これらのコーパスを読者の対象年齢を基準に4グループに分類した．

帰納的学習により，対象に動的に適応できるかどうかを調べるための実験を2通り行った．どのような性質をもつ統語規則を獲得できるのかを調べる実験Iと，入力文の種類によって学習にどのような影響が現れるかを調べるための実験IIである．実験Iでは，文献[27]を30回繰り返して入力した．繰り返して入力

表1 実験で使用したコーパス
Table 1 Corpora used in the experiments.

コーパスの内容	文数	グループ番号
外国人の日本語学習用テキスト[27]	860	1
科学の絵本[28]	340	2
小学1年国語[29]	132	3
小学2年国語[30]	300	3
小学3年国語[31]	256	3
中学1年国語[32]	64	4
中学公民[33]	1,508	4

する理由は，幼児の言語獲得過程において繰返しが有効である[34]からである．実験IIは，表1を文献[27]から文献[33]の順に1回だけ通して入力した．実験IIは実験Iと比較して，多様な言語表現が入力されることになる．どちらの実験においても，事前に備えている統語規則が学習に影響を与えるのを避けるために，統語規則辞書が空の状態から実験を行った．

5.1 評価基準

本手法で獲得される規則は，統語解析のための規則である．したがって，我々の評価は解析結果Forestに基づいて行った．図2で示しているように，形態素解析結果GroveはForestの底辺として表現されている．したがって，Forestを評価するということは形態素解析と構文解析の両方の結果を評価していることになる．一般には，解析結果が正しいかどうかのみを評価基準としているが，本手法により獲得される統語規則は，すべて解析に適用できることを保証されていないことから，解析の成功と解析結果の正解の2点を数値に基づいて評価することとした．Forestの評価基準として，解析成功率（Rate of Succesful Forest，RSF）と解析正解率（Rate of Correct Forest，RCF）の二つを設けた．

$$RSF = \frac{\text{解析が成功した入力文数}}{\text{入力文の総数}} \quad (5)$$

$$RCF = \frac{\text{解析が正解である入力文数}}{\text{解析が成功した入力文数}} \quad (6)$$

解析成功は，入力文を，最上層FloraのPlantの数が1であるForestで表現できたことを示す．3.3で述べたように，未知語が存在するForestは不完全である．しかしながら，最上層FloraのPlantの数が1であるならば，その未知語の品詞に相当する統語範疇が推測されていることになる．ゆえに，解析の成功基準に未知語の有無を含めない．解析正解は，解析成功であるForestの中で，すべての文節以上の単位において，文節区切りと係り受け関係に誤りがないことを示す．日本語構文解析の主流である係り受け解析（例えば，[35]）において，文節単位で解析が行われていることから，文節よりも小さい単位での文節区切りと係り受け関係を調べる必要がないと判断した．本手法の形態素解析では，入力文を分割した後，分割された文字列を対応する統語範疇に変換している．しかしながら，統語範疇は番号で識別されており，本論文では，「名詞」や「動詞」などの品詞が，何番の統語範疇に対応するのかという問題については，まだ扱っていない．したがって，今回の評価においては，品詞化の正否を含めていない．また，この誤りがないかどうかの判定は，第1筆者が行った．

5.2 実験Iの結果と考察

実験IのRSFを表2と図4に示す．図4から，入力回数が増えるに従って，RSFが全体的に上昇していることがわかる．最終的に30回繰り返した時点で，RSFは85.3%になった．このことは，本手法が，統語規則辞書が空の状態からタグなしコーパスを入力として，入力全体を一つの統語範疇で表すことができる統語規則を獲得したことを示している．しかしながら，17回目から18回目に移るときにわずかながら減少している．17回目で解析が成功していながら，18回目で解析が失敗した入力例として，

- たなからコップをとりました

があった．17回目では「たな/から」と形態素解析を行い，両者を統合する統合規則に従って成功させていた．18回目では「たなか/ら」と形態素解析を行ったため，両者を統合する統合規則が存在せず失敗となった．これは，17回目から18回目の間に，

- たなかさんは…（田中さんは…）

などの入力文から，認知単位「たなか」をもつ認知規則が新たに獲得されたことが原因と思われる．「たな」と「たなか」に対応する統語範疇は同じDominantをもっており，適用する範囲も交差していないため，4.1で述べた認知規則を適用するときの基準に従うと，3

表2 実験Iの RSF.a
Table 2 Rate of successful Forest in the experiment I.a.

入力回数	1	10	17	20	30
RSF	11.6%	56.7%	76.9%	75.7%	85.3%

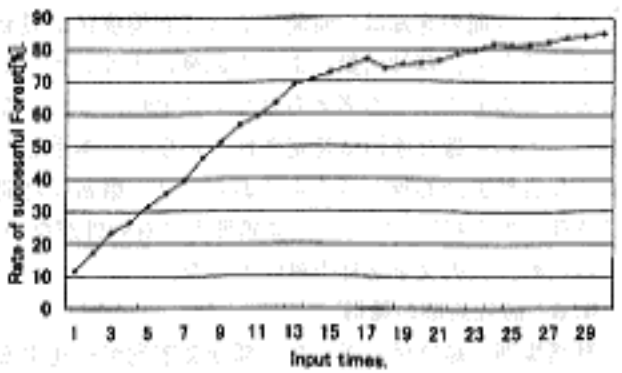

図4 実験Iの RSF.b
Fig. 4 Rate of successful Forest in the experiment I.b.

文字の「たなか」が優先して適用される．仮に，**4.1** の基準をすべて無視し，すべての判断をゆう度評価に委ねた場合を考える．**4.1** の Dominant の定義から，「ら」に対応する統語範疇の Dominant は 1 であり，「たな」，「から」，「たなか」に対応する統語範疇の Dominant は 2 である．**4.3** の式 (3) に従って，PDG を求めると，「たな/から」の場合，$\frac{2\times2+2\times2}{2}=4$ となる．「たなか/ら」の場合，$\frac{2\times3+1\times1}{2}=3.5$ となる．したがって，「たな/から」とした正解の方が高い PDG が得られる．しかしながら，4.1 の基準を無視して形態素解析を行うことは，すべてを文字にまで分解した結果が得られるまで解析を行うことになる．これでは，認知単位や Dominant を使用する有効性がなくなり，あいまい性の増加につながる．この調整をどうするかは，今後の課題である．

次に，RCF について調べる．RCF は，30 回目の入力に対する Forest をもとに算出した．解析正解である文の例を下に挙げる．

- ((なべは)((たなに)(ありますか)))
- ((かにには)((はさみが)(あります)))
- ((これが)((わたしの)(左耳です)))
- (((この)(ドアは))(とじています))

" (" と ") " は，正しい認知単位の区切りと係り受け関係を表す．認知単位 α，β を (α)，(β) と表現し，α と β が係り受け関係にあることを $\langle(\alpha)\ (\beta)\rangle$ と表現する．また，誤りのない部分の文節より小さい認知単位と係り受け関係の括弧は省略している．次に，解析正解ではない解析成功の文を例示する．誤りと判断した部分の係り受け関係を " 《" と " 》 " で表現した．

- ((((ど)《(の)(かん)》)(が))(コーヒーですか))
- ((これは)((はり)《(の)(目です)》))
- ((山下)《(さんは)((ありがとう)《(と)(いっています)》)》)
- ((ひとつの)((コップ)(《(は)(右手)》《(に)(あります)》)))

成功した Forest は 734 あり，その内，正解の Forest は 314 であった．その結果，RCF は $\frac{314}{734}\approx 42.8\%$ となった．条件が異なるため，単純な比較はできないが，参考までに，括弧付きコーパスから確率文脈自由文法規則を獲得する文献 [5] の実験結果を述べる．文献 [5] では，受理率が約 92%，生成確率 1 位の文の正解率が約 8.5%，生成確率 30 位までの文の正解率が約 29.1%となった．ここでの「受理」とはパーザが解析に成功して 1 個以上の解析木を出力できたことを示す．「正解」とは，解析結果のすべての係り受け関係がコーパス中のどの構文構造とも交差していないことを示す．本手法との比較においては，受理率は RSF に相当し，文の正解率は RCF に相当すると考えられる．本手法と比較した結果，RSF でわずかに下回るが，生成確率 30 位までの結果と比較しても，RCF で大きく上回っていることが確認できる．これは，本手法の有効性を示していると考えられる．しかしながら，一般的な解析結果として見た場合，これは決して十分な数値ではない．Forest を調べた結果，

- テーブル/にあります
- ここ/にあります

などの単語境界の誤りが目立った．これは，同じコーパスを繰り返して入力したため，同じ表現が何度も出現し，人間から見て不自然であっても，その表現専用に特化した統語規則が獲得されたことが原因と考えられる．「に」の直後に「あ」が存在する確率と，「あ」の直前に「に」が存在する確率をグループごとに調べた．その結果を表 3 と表 4 に示す．第 1 グループである文献 [27] が，他の文献と比較して，「に」と「あ」の連続して共起する確率が最低でも 5 倍以上高いことがわかる．文献 [27] における RCF 低下の原因は，「にあります」を一つの認知単位とみなす方が処理効率が良いとシステムが判断したことにある．これは，対象に動的に適応するという能力が過剰に反映された結果である．RCF を高めるためには，入力文への適応を制御する必要があると考えられる．これは，今後の課題である．また，係り受け関係における誤りとして，

- おと《(この)(ひと)》…

などがあった．「この」は「ひと」にかかるという規則自体は間違いではない．しかしながら，「おとこのひと」という文脈においては間違いである．単純な文脈自由文法で，そのような文脈を考慮することはできない．そのため，文脈自由文法を獲得対象とする限り，この

表 3 「に」の直後に「あ」が存在する確率
Table 3 Probability that "あ" adjoins after "に."

グループ番号	1	2	3	4
存在確率	24.5%	0.6%	2.0%	2.2%

表 4 「あ」の直前に「に」が存在する確率
Table 4 Probability that "に" adjoins before "あ."

グループ番号	1	2	3	4
存在確率	32.6%	0.9%	4.3%	6.1%

誤りに対して根本的に解決することはできない．したがって，獲得する頑語規則として，文脈自由文法よりも強力な文法を獲得の対象にする必要がある．

また，30回目の入力文860文を処理するためにかかった時間は，CPUがAlpha21164A-500 MHz，メインメモリが512MBの計算機を使用して168分であった．

5.3 実験IIの結果と考察

実験IIのRSFを表5と図5に示す．実験IIでは，実験Iと違い十分な学習が行われていない．まず，原因として考えられるのが，入力文数の違いである．実験Iの延べ入力文数は $860 \times 30 = 25,800$ であるのに対して，実験IIの延べ入力文数は $860+340+132+300+256+64+1,508 = 3,460$ である．しかしながら，実験1の $\frac{3,460}{860} \approx 4$ 回目のRSFと比較しても十分な学習が行われていない．そこで，表1のグループごとに，入力するコーパスの統語的特徴による影響を考える．図5から，特に，第1グループから第2グループに移るときと，第3グループから第4グループに移るときに大幅なRSFの低下が見られる．各グループごとの統語的特徴を調べるために，平均文字数（Average Number of Character，ANC）と字種別出現確率（Probanility of Appearance of each Kind of Character，PAKC）を調べた．

$$ANC = \frac{\text{すべての文字の出現回数}}{\text{入力文の総数}} \tag{7}$$

$$PAKC = \frac{\text{ある字種の総出現回数}}{\text{総文字数}} \tag{8}$$

字種は，平仮名，片仮名，漢字，その他（句読点や記号など）に分類して調べた．ANCを表6，PAKCを表7に示す．ANCは，第1グループと第2グループの間，第3グループと第4グループの間で，2倍前後の文字数に増加する．PAKCは，第3グループと第4グループの間で，平仮名の出現確率が減少し，代わって漢字の出現確率が大きく増加している．以上は，RSFの低下する時期と一致している．内容的には大きく異なる第2グループと第3グループは，ANCとPAKCの点では大きく異ならない．学習に影響する他の要因が存在しないか調べるために，第2グループから第3グループにかけての1文当りの平均上昇率と比較する．図4を見ると，17回目の入力を境として上昇率がわずかに変化している．そこで，全体を通しての1文当りの上昇率と，1回目から17回目の間での1文当りの上昇率を求めて比較した．第2グループから第3グループの平均上昇率は，$\frac{4.3-0.3}{340+132+300+200} \approx 3.9 \times 10^{-3}$ である．実験Iの全体の平均上昇率は，$\frac{85.3-11.6}{860 \times 30} \approx 2.9 \times 10^{-3}$ であり，1回目から17回目の間での平均上昇率は，$\frac{76.9-11.6}{860 \times 17} \approx 4.5 \times 10^{-3}$ である．したがって，1文当りの上昇率には，それほど大きな差がないと判断できる．以上のことから，文字の種類と文字数における環境の変化が，学習に影響を与えていると考えられる．文字の種類と文字数が大きく変動しないように入力を行うことで，効率良く学習が行われると考えられる．

表5 実験IIのRSF.a
Table 5 Rate of successful Forest in the experiment II.a.

コーパスの内容	RSF
外国人の日本語学習用テキスト[27]	11.8%
科学の絵本[28]	0.3%
小学1年国語[29]	1.5%
小学2年国語[30]	0.7%
小学3年国語[31]	4.3%
中学1年国語[32]	0.0%
中学公民[33]	0.1%

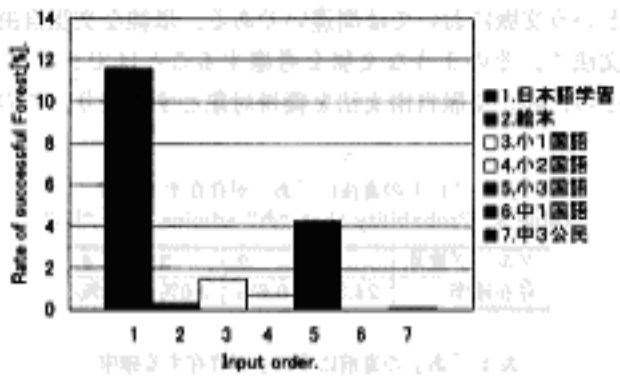

図5 実験IIのRSF.b
Fig. 5 Rate of successful Forest in the experiment II.b.

表6 平均文字数
Table 6 Average number of character.

グループ番号	1	2	3	4
ANC	12.6	25.8	23.6	41.5

表7 字種別出現確率
Table 7 Probability of appearance of each kind of character.

字種	1	2	3	4
平仮名	82.0%	79.1%	79.8%	51.5%
片仮名	4.8%	2.2%	2.0%	2.4%
漢字	3.0%	0.3%	8.9%	38.5%
その他	10.2%	18.4%	9.3%	7.8%

6. むすび

本論文では，入力文に対して，形態素解析と構文解析を行い，その解析結果を構成する要素同士の位置と順序関係に着目し，類推と統計的基準に基づく帰納的学習を用いて，統語規則を自動的に獲得する手法を提案した．その結果，本手法を実装したシステムは，文の種類を限定せずに，分かち書きされず，いかなるタグも付けられていない文を入力として統語規則辞書がすべて空の状態から統語規則を獲得できるという頑健なシステムとなった．また，そのシステムの性能に対する評価実験を以下のように行った．まず，統語規則辞書が空の状態から，文献[27]を30回繰り返して入力した結果，85.3%のRSFが得られた．また，解析成功事例の42.8%のRCFを得られる統語規則が獲得され，本手法の有効性を確認することができた．しかしながら，半数以上の誤りがあった．その誤りの原因は，入力文に適応しすぎたことによるものであった．これを解決するためには，入力文への適応を制御する必要がある．次に，7種類のコーパスを連続して入力し，入力するコーパスによって本手法の学習にどのような影響があるのかを調査した．その結果，文字の種類と文字数が学習に影響を与え，効率の良い学習には，文字の種類と文字数に大きな差のない入力が適していることが明らかになった．今回は，日本語を対象としたが，タグなしコーパスを入力として，統語規則辞書が空の状態から処理することができる本手法は，他の言語に対しても適用できる可能性を秘めている．今後の課題として，解析結果の選択における形態素解析とゆう度評価の間の調整，正解率を高めるような入力への適応の制御，文脈自由文法に代わる文脈などの情報を扱える強力な統語規則の獲得などが挙げられる．

謝辞　本研究の一部は，文部省科学研究費補助金（No.10680367）により行われた．

文　献

[1] 長尾　真（編），自然言語処理，岩波ソフトウェア科学15，岩波書店，東京，1996．

[2] 田中穂積，自然言語解析の基礎，産業図書，東京，1989．

[3] 荒木健治，桃内佳雄，中村順一，"帰納的学習による語の獲得および確実性を用いた語の認識，" 信学論（D-II），vol.J75-D-II，no.7，pp.1213-1231，July 1992．

[4] 森　信介，長尾　真，"統計によるタグ付きコーパスからの統語規則の獲得，" 情処学NL研報，vol.110，no.12，pp.79-86，Nov. 1995．

[5] 白井清昭，徳永健伸，田中穂積，"括弧付きコーパスからの日本語確率文脈自由文法の自動抽出，" 自然言語処理，vol.4，no.1，pp.125-146，Jan. 1997．

[6] 西岡山滋之，宇津呂武仁，松本裕治，"コーパスからの日本語従属節係り受け選好情報の抽出，" 情処学NL研報，vol.126，no.5，pp.31-38，July 1998．

[7] K. Lari and S.J. Young, "The estimation of stochastic context-free grammars using the inside-outside algorithm," Computer Speech and Language, vol.4, pp.35-56, 1990.

[8] M. Kiyono and J. Tsujii, "Hypothesis selection in grammar acquisition," Proc. 15th COLING, vol.2, pp.837-841, 1994.

[9] M. Kiyono and J. Tsujii, "Combination of symbolic and statistical approaches for grammatical knowledge acquisition," Proc. 4th Conference on Applied Natural Language Processing, pp.72-77, Oct. 1994.

[10] 鍋見美貴子，中島秀之，松原　仁，"一般学習機構を用いた言語獲得の計算機モデル，" 認知科学の発展，vol.5，日本認知科学会（編），pp.143-185，講談社，東京，1992．

[11] 田中春美，樋口時弘，家村睦夫，五十嵐康男，下宮忠雄，田中幸子，入門ことばの科学，大修館書店，東京，1994．

[12] 乾健太郎，脇川清和，"品詞タグつきコーパスにおける品詞体系の変換，" 情処学NL研報，vol.132，no.12，pp.87-94，July 1999．

[13] M.K. Tanenhaus, J.M. Leiman, and M.S. Seidenberg, "Evidence for multiple stages in the processing of ambiguous words in syntactic contexts," J. Verbal Learning and Verbal Behavior, vol.18, no.4, pp.427-440, Aug. 1979.

[14] D.A. Swinney, "Lexical access during sentence comprehension: (Re)consideration of context effects," J. Verbal Learning and Verbal Behavior, vol.18, no.6, pp.645-659, Dec. 1979.

[15] 横田和章，亀田弘之，藤崎博也，"日本語の文法および未知の認知単位の自動獲得のための一方法，" 自然言語処理，vol.3，no.4，pp.115-128，Oct. 1996．

[16] 小川泰弘，ムフタル マフスット，外山勝彦，稲垣康善，"派生文法による日本語形態素解析，" 情処学論，vol.40，no.3，pp.1080-1090，March 1999．

[17] 波多野誼余夫，"Connectionist infantsは統語規則を獲得しうるか，" 心理学評論，vol.40，no.3，pp.319-327，1997．

[18] ノーム チョムスキー，言語と知識—マナグア講義録（言語学編），田中行則，都司隆男（訳），産業図書，東京，1989．

[19] 松沢哲郎，チンパンジーから見た世界，波多野誼余夫（道案内），東京大学出版会，東京，1991．

[20] Y. Lepage and S. Ando, "Saussurian analogy: A theoretical account and its application," Proc. COLING-96, vol.2, pp.717-722, Aug. 1996.

[21] J.R. Saffran, R.N. Aslin, and E.L. Newport, "Statistical learning by 8-month-old infants," Science, vol.274, pp.1926-1928, Dec. 1996.

[22] 荒木健治，桃内佳雄，永田昂一，"多段階分割法によるべた書き日本語文のかな漢字変換，" 情処学論，vol.28，no.4，pp.412-421，Dec. 1987．

[23] D.D. McDonald, "An efficient chart-based algorithm for partial-parsing of unrestricted texts," Proc.

3rd Conf. on Applied Natural Language Processing, pp.193–200, March–April 1992.
[24] 渋木英潔，荒木健治，栃内香次，"平文を用いた統語規則の自動獲得，" 信学技報，TL99-8, July 1999.
[25] C.D. Manning, "Automatic acquisition of a large subcategorization dictionary from corpora," Proc. 31st Annual Meeting of the Association for Computational Linguistics, pp.235–242, June 1993.
[26] 御領 謙，菊地 正，江草浩幸，最新認知心理学への招待，サイエンス社，東京，1993.
[27] 片桐ユズル，はじめてのにほんご，大修館書店，東京，1990.
[28] デニス＝リグリー，はじめてのかがく 1–15，中川 宏（訳），偕成社，東京，1987.
[29] 木下順二，柴田 武，外山滋比古，今西祐行，ほか30名，こくご1上，教育出版，東京，1998.
[30] 木下順二，柴田 武，外山滋比古，今西祐行，ほか30名，国語2上，教育出版，東京，1998.
[31] 木下順二，柴田 武，外山滋比古，今西祐行，ほか30名，国語3上，教育出版，東京，1998.
[32] 木下順二，松村 明，加藤周一，ほか32名，中学国語1，教育出版，東京，1998.
[33] 田邊 裕，吉田 孝，阪上順夫，ほか37名，新編新しい社会 公民，東京書籍，東京，1998.
[34] ジョージ バターワース，マーガレット ハリス，発達心理学の基本を学ぶ，村井潤一（監訳），小山 正，神土陽子，松下（共訳），ミネルヴァ書房，東京，1997.
[35] 春野雅彦，白井 諭，大山芳史，"決定木を用いた日本語係り受け解析，" 情処学論，vol.39, no.12, pp.3177–3186, Dec. 1998.

（平成11年11月26日受付，12年5月24日再受付）

渋木 英潔 （学生員）

平9小樽商科大・商・商業教員養成課程卒，平11同大大学院修士課程了，現在，北大大学院博士課程に在学中．自然言語処理に関する研究に従事．情報処理学会，言語処理学会各会員．

荒木 健治 （正員）

昭57北大・工・電子卒，昭63同大大学院博士課程了．工博．同年，北海学園大学工学部電子情報工学科助手，平1同講師，平3同助教授，平10同教授．現在，北大・工・電子情報工学専攻助教授．自然言語の機械学習と機械翻訳に関する研究に従事．情報処理学会，日本認知科学会，人工知能学会，言語処理学会，IEEE，ACL，AAAI各会員．

栃内 香次 （正員）

昭37北大・工・電気卒．昭39同大大学院工学研究科電気工学専攻修士課程了．現在，北大・工・電子情報工学専攻教授．主として音声情報処理，自然言語処理の研究に従事．工博．情報処理学会，日本音響学会各会員．