修　士　論　文

題　目　統計によるコーパスからの統語規則の獲得

指導教官　長尾　眞　教授

京都大学大学院工学研究科 電気工学第二専攻

氏 名　森　信介

1995年2月10日

目 次

第 1 章　序論 **1**

第 2 章　文法学習と言語統計 **4**

2.1　形式言語と文法学習 4

2.1.1　形式言語 4

2.1.2　形式言語の帰納学習 7

2.2　大規模言語コーパス 8

2.2.1　付加情報 9

2.2.2　コーパスの作成 9

2.3　自然言語の文法学習 10

2.3.1　自然言語の生成規則 10

2.3.2　文法学習 11

2.4　統計的手法 15

2.4.1　出現頻度 16

2.4.2　出現確率 16

2.4.3　確率分布 16

2.4.4　エントロピー 17

2.4.5　分散 18

2.4.6　分布分析 19

第 3 章　統計による統語規則の抽出 **20**

3.1　統計による統語規則の抽出 20

3.1.1　生成規則と文形式の関係 20

3.1.2　品詞列の独立性 20

3.1.3　品詞列の類似性 23

3.2　抽出と適用の反復 25

3.2.1　n-gram 統計の計算 26

3.2.2 生成規則の抽出 26
3.2.3 コーパスへの生成規則の適用 29
3.2.4 反復による新たな規則の獲得 31

第 4 章 実験結果と考察 **32**
4.1 コーパスのデータ 32
4.2 実験結果と評価 32
4.2.1 評価基準 33
4.2.2 実験の条件 34
4.2.3 仮定を実現する手法の妥当性 35
4.2.4 解析結果の評価 36
4.2.5 生成規則の評価 37
4.3 考察 38

第 5 章 結論 **39**

謝 辞 **41**

参 考 文 献 **42**

付録 A 数式の変形 **65**
A.1 分散 65
A.2 エントロピーの最大値 65
A.3 分散の最小値 66

付録 B 解析結果の例 **67**
B.1 方法 1 67
B.2 方法 2 74
B.3 方法 3 80

第 1 章 序論

自然言語を研究対象とする言語学は、言語現象の曖昧性から伝統的に定性的な議論に終始する傾向があった。しかし、1948 年にシャノン (C. E. Shannon) は情報理論を確立し、自然言語をマルコフ過程としてとらえる立場を提案した [1)]。これは、それ以前の考え方とは全く異なり、次のような極めて数理的なモデルを仮定している。自然言語をマルコフ過程とみなし、その出力から自然言語の抽象的構造を数学的に記述する。具体的には、ある自然言語に属する文を大量に集めたコーパス中に、その言語を記述するために用いられる文字がどのように出現するかを観測し、その結果に基づいてモデルのパラメーターを決定する。換言すると、ある n 個の文字の組み合せ (n グラム) の頻度を計算し、この結果を用いて可能な限り近い性質の出力を持つマルコフモデルを構成し、これを自然言語を近似するモデルとするのである。このとき、注目する文字列の長さ n が大きくなる程よい近似となることが示せる。

しかし、当時は大規模な機械可読のコーパスもなく、大量のデータを高速に処理する計算機も存在しなかったため、十分な実験とその結果に対する考察を行なうことができなかった。例えば対象言語を英語とし、それを記述する文字として、アルファベット、空白、終始符、疑問符、感嘆符の 30 文字を仮定すると、注目する文字列の長さを $n = 5$ としても $30^5 = 24{,}300{,}000$ 通りの文字列に対して出現頻度を計算する必要があり、これには膨大な記憶領域と計算時間を要すると考えられていた。このことに加えて、信頼性の高い統計結果を得るためには非常に大規模なコーパスが必要となる。このような理由から、注目する文字列の長さとして $n = 3$ 程度が限度であった。

近年計算機技術が発達し、言語を数理的に分析する計算言語学の分野で待ち望まれていた大規模な機械可読な言語データが次々と現れてきている。例えば、本研究で用いた Penn Treebank [2) 3) 4)] には、文章の他に各単語の品詞情報や各文の統語構造情報などが付加されている。一方、情報を処理する計算機技術もハードウェア、ソフトウェアの両方において進歩し、現存するコーパスに対して任意の長さの文字列の出頻度を現実的な時間で計算することが可能になった [5)]。以上に述べた事実は、シャノンが提案した文字を対象とした統計のみならず、品詞や

統語構造を対象としたより複雑な統計を行なうことが可能になったことを意味する。

ところで、シャノンが提案した自然言語のモデルは、一般的な自然言語観とは全く異なり、句や構造というような側面を持たない単純なものであった。それに対し、チョムスキー (N. Chomsky) は、自然言語の再帰的性質や呼応表現の存在を根拠に句構造文法による自然言語のモデル化を提案した[6]。チョムスキーが提案した自然言語の3つのモデルは形式言語理論として整理され、それらの間に階層的な関係があることから、現在チョムスキーの階層として知られている。

句構造文法が提案されてからも、さまざまな文法が自然言語を記述する目的で提案されている。しかし、社会的に要求されている自然言語処理を応用したさまざまな実用システムを設計し実装するという視点からは、ある文法を計算機上に記述し、これを用いて実際に自然言語処理を行なうことが重要である。このように、工学的視点から見れば、自然言語を記述する文法は完全性のみならずその実装の可能性を考慮する必要があり、これは完全性よりも優先される課題となることもある。ある文法を計算機上に実装することは人間にとって困難な作業であり、ひとたび文法が記述されたとしても、その記述された文法を用いた解析の結果に基づく修正が不可避となる。しかし、文法のある部分を他の部分との整合性に注意しながら修正していくことは、規則数の増加に伴い非現実的となる。以上に述べたことから、計算機による文法規則の獲得や学習が必要であると帰結できる。形式言語の文法学習についての研究は、すでにいくつかなされており[7] [8] [9] [10]、これらの結果は自然言語の文法学習にも応用することができると考えられるが、与えられる情報が不正確であることなどの自然言語特有の問題があり、これらを直接利用することはできない。しかし、大規模な計算機可読なコーパスが出現し、これを処理するための計算機環境も整ってきているので、統計を用いることが可能である。このような背景を考慮すると、計算機による自然言語の文法規則の獲得や学習は十分可能であり、研究の必要がある。また、言語データを基にして獲得された文法規則は、より良く自然言語を説明すると考えられ、主に人間の言語直観による従来の文法を再検討する一助になると思われる。

以上に述べた考察に基づき、我々は Penn Treebank に含まれる品詞情報が付加された英語コーパス Wall Street Journal に対し品詞単位の統計を行ない、その結果を用いた文法規則の獲得について研究を行なった。自然言語のモデルとして、正規文法および文脈自由文法を用いた。我々が提案する手法は以下の3つの仮定

に基づいている。1) 生成規則の右辺に現れる記号列は環境から独立して高い頻度で現れる。2) 同じ非終端記号から直接導出される品詞列は類似した環境を持つ。3) コーパスの大きさを最も減少させる生成規則が適切である。類似の研究として Fernando Pereira と Yves Schabes [11) 12)] は確率文脈自由文法をモデルとし、すべての可能なチョムスキー標準形の書き換え規則の確率をコーパスに付加された統語構造から計算するアルゴリズムを提案している。また、Eric Brill [13)] は、まず英語が右分岐の言語であることを根拠にコーパス中のすべての文に対し右線形の統語構造を与え、あらかじめ与えられた正しい統語構造と比較し誤りを最も良く修正する構造変形規則を繰り返し獲得していく手法を提案している。これらの試みに対して、我々は統語構造情報を用いない、品詞情報のみからの文法規則の獲得の研究を行なった。これは、品詞情報の付加に関しては高い精度のアルゴリズム[14) 15) 16) 17)] が提案されていることから修正など後処理のコストは低いと考えられるが、統語構造情報の付加に関しては高い精度のアルゴリズムが存在しない [18)] ことから後処理のコストが品詞情報付加に比べて非常に高くなるためである。本論文では、上述した3つの仮定に基づきコーパスを品詞列と見なして生成規則を獲得する手法を提案し、この手法による統語規則の獲得の実験結果を評価する。

本論文の構成は以下の通りである。まず第2章で、本論文で扱う問題の定義とその解決のために我々が用いた統計的手法の基礎について説明する。次に第3章で、これらの基本的手法を用いた文法獲得の具体的手法について説明する。さらに第4章では、この手法を実現し実験を行なった結果について述べる。最後に、第5章で本研究の結論を述べる。

第 2 章 文法学習と言語統計

本研究では、自然言語のモデルとして形式言語を用いた。この理由は、形式言語理論では言語をアルファベットの列とみなし、その構造を表層に現れる文字情報のみを利用して記述することである。具体的には、比較的単純で性質が良く知られてる文脈自由文法 (context free grammar, CFG) 及び正規文法 (regular grammar, RG) を用いた。これは、ある複雑な文法を仮定するということで自然言語の性質を根拠なく規定することを最小限にとどめるためである。

この章では、まず形式言語とその文法の学習についての簡単な説明を行なう。また、これと並行して、本論文で用いる記号の定義を与える * 。次に、自然言語の文法学習に必要となる大規模言語コーパスについて説明する。さらに、自然言語の文法学習について自然言語と形式言語の差異という点から説明し、関連研究について述べる。最後に、本論文における問題の定義とその解決法として我々が提案する統計的手法の基礎について説明する。

2.1 形式言語と文法学習

チョムスキーが自然言語を定式化する目的で提唱した句構造文法は、形式言語理論として整理され、発展してきた。この節では、最も制限が緩い句構造文法 (phrase structure grammar, PSG) と、基本的な 3 つの文法、すなわち、文脈依存文法 (context sensitive grammar, CSG) 、文脈自由文法、正規文法について簡単に説明する。次に形式言語の帰納学習について述べる。

2.1.1 形式言語

自然言語の文は、その言語を表すための文字の連続であるとみなすことができる ** 。例えば、英語の文はアルファベットと空白を含めたいくつかの記号で構成される文字の連続とみなせる。このとき、文法はその文字の並び方に関する規則

* 記号の定義は John E. Hopcroft と Jeffrey D. Ulman [19] に準拠している。

** 音声などを媒体としている場合はこの限りではない。

である。このような考えに基づいて自然言語を定式化するために導入された文字とその並び方に関する概念が、コンパイラなど自然言語の記述以外の応用などにともなって発展し、形式言語理論として体系化された。

形式言語理論では、記号の空でない集合をアルファベット (alphabet) と呼ぶ。記号を一列に並べてできる列を記号列 (string) 呼ぶ。例えば、a,b,c が記号のとき $abcb$ は記号列である。記号列 ω を構成する記号の数を ω の長さ (length) と呼び、$|\omega|$ で表す。0 個の記号からなる記号列を空列と呼び、ε で表す ($|\varepsilon| = 0$) 。アルファベットに属する記号の列のある集合を言語 (language, くわしくは formal language) と呼ぶ。以下本論文ではアルファベットを主に T で表す。あるアルファベット T の上のすべての記号列の集合も一つの言語である。この言語を T^* で表す。例えば、$T_1 = \{a,b\}$ とすると $T_1^* = \{\varepsilon, a, b, aa, ab, ba, bb, aaa, \cdots\}$ である。自然言語の例として英語をとりあげると、アルファベット T_e は通常のアルファベットに加えて空白やその他のいくつかの記号を含み、英文はアルファベットの連続であるから $a_1 a_2 \cdots a_n$ $(a_k \in T_e, k = 1,2,\ldots,n)$ で表される記号列の中のある規則を満たすものである。この規則は一般に文法と呼ばれるが、形式言語理論でも同様に文法 (grammar) と呼ぶ。英語 L_e はある文法規則を満たす文の集合であるから T_e^* の部分集合となる ($L_e \in T_e^*$)。

文法は通常 $G = (N,T,P,S)$ の 4 項組で表わされる。ここで、N は非終端記号 (nonterminal symbol) の有限集合、T はアルファベットであり、その要素を非終端記号との比較から終端記号 (terminal symbol) と呼ぶ。N と T は共通の要素を持たないと仮定する ($N \cap T = \phi$) 。P は生成規則 (production, 書き換え規則ともいう) の有限集合であり、生成規則は $\alpha, \beta \in (N \cup T)^*$ として $\alpha \to \beta$ という形をしている。S は開始記号 (start symbol) と呼ばれる特別な非終端記号である。

さらに、文法 G における $(N \cup T)^*$ の 2 つの要素の関係 $\Rightarrow$ および $\overset{*}{\Rightarrow}$ を定義する。もし $\alpha \to \beta \in P$ で、$\gamma, \delta \in (N \cup T)^*$ であれば、$\gamma\alpha\delta \Rightarrow \gamma\beta\delta$ である。このとき、文法 G において $\gamma\alpha\delta$ から $\gamma\beta\delta$ が直接に導出 (directly derive) されるという。さらに、$\alpha_1, \alpha_2, \cdots, \alpha_n \in (N \cup T)^*$ として以下の関係が成り立つとしよう。

$$\alpha_1 \Rightarrow \alpha_2,\ \alpha_2 \Rightarrow \alpha_3,\ \cdots,\ \alpha_{n-1} \Rightarrow \alpha_n$$

このとき、文法 G において α_1 から α_n が導出 (derive) されるといい、$\alpha_1 \overset{*}{\Rightarrow} \alpha_n$ と書く。

G によって生成される言語 (language) $L(G)$ とは、$\{\omega \mid \omega \in T^*, S \overset{*}{\Rightarrow} \omega\}$ のこと

である。終端記号と非終端記号の列 $\alpha \in (N \cup T)^*$ は、$S \overset{*}{\Rightarrow} \alpha$ であるとき文形式 (sentential form) と呼ばれる。

以上のように定義される形式言語は、文法 G の書き換え規則 P の制限によって、クラスを構成する。以下では、まずこのクラスの中で重要な 4 つについて以下で説明し、次にクラス間の関係について述る。

句構造文法

句構造文法 (phrase structure grammar, PSG) と呼ばれる文法は、生成規則 (P の要素) が次の形式のものである。

$$\alpha \to \beta \ \ (\alpha \in (N \cup T)^* N (N \cup T)^*,\ \beta \in (N \cup T)^*)$$

この文法で生成される言語を句構造言語 (phrase structure language, PSL) と呼ぶ。

文脈依存文法

句構造文法 $G = (N, T, P, S)$ の生成規則が以下のように制限されているとき

$$\beta A \gamma \to \beta \alpha \gamma \ \ (\alpha, \beta, \gamma \in (N \cup T)^*,\ A \in N,\ \alpha \neq \varepsilon)$$

G は文脈依存文法 (context sensitive grammar, CSG) と呼ばれる。この文法で生成される言語を文脈依存言語 (context sensitive language, CSL) と呼ぶ。

文脈自由文法

句構造文法 $G = (N, T, P, S)$ の生成規則が以下のように制限されているとき

$$A \to \alpha \ \ (A \in N,\ \alpha \in (N \cup T)^*)$$

G は文脈自由文法 (context free grammar, CFG) と呼ばれる。この文法で生成される言語を文脈自由言語 (context free language, CFL) と呼ぶ。

正規文法

句構造文法 $G = (N, T, P, S)$ の生成規則が以下のように制限されているとき

$$A \to aB \ \text{ あるいは } \ A \to a \ \ (A,\ B \in N,\ a \in T)$$

G は正規文法 (regular grammar, RG) と呼ばれる。この文法で生成される言語を正規言語 (regular language, RL) と呼ぶ。

言語族とその間の階層関係

上述の文法には表現能力に差があり、それぞれの文法で表すことができる言語の間には、包含関係がある。これを表現するために、言語族という概念を導入する。言語族とは、ある条件を満たす言語の集合である。例えば、$L_1, L_2, \ldots, L_n$ を句構造言語とすると、句構造言語族は以下のように定義できる。

$$\mathcal{PSL} = \{L_1, L_2, \ldots, L_n\}$$

この定義に加えて、文脈依存言語族、文脈自由言語族、正規言語族をそれぞれ $\mathcal{CSL}, \mathcal{CFL}, \mathcal{RL}$ とすると、言語間の包含関係は以下のように表すことができる。

$$\mathcal{PSL} \supset \mathcal{CSL} \supset \mathcal{CFL} \supset \mathcal{RL} \tag{2.1}$$

これをチョムスキー階層と呼ぶ。

2.1.2 形式言語の帰納学習

形式言語の帰納学習とは帰納推論に基づく学習のことで、以下のように定義される。

> ある言語に関する有限の具体例から、その言語を規定している文法体系を学習すること

ここでいう具体例には以下の二種類がある。

- **正の例** (C^+) その言語に属す文。 $C^+ = L(G)$
- **負の例** (C^-) その言語に属さない文。 $C^- = T^* - L(G)$

ここで、$L(G)$ は問題となっている言語を、T はそのアルファベットを表す。形式言語の学習では、アルファベットは正の例を表すために必要な文字であり、開始記号と非終端記号は任意に選んでも一般性を失わないので、生成規則を学習することが目標となる。例えば、

$$C^+ = \{ab, abc, abd\},\ L(G) \in \mathcal{CFG}$$

という情報から、生成規則の要素として以下を学習する。

$$S \to AB \qquad A \to a \qquad B \to C \qquad B \to D$$
$$B \to E \qquad C \to b \qquad D \to bc \qquad E \to bd$$

ここで、A, B, C, D, E は非終端記号である $(A, B, C, D, E \in N)$。

形式言語の学習では、上に述べた例のように対象とする言語族を限定する。この理由の一つとして以下のことが挙げられる。

> 最も一般的な句構造文法を学習し、その生成規則の特徴から結果として言語族が規定されることが理想的であるが、句構造文法の学習は現段階では不可能である。

しかし、このように言語族を生成規則に対する制限が最も大きい正則言語の族に限定しても、一般的に起こり得る正の例だけからはうまく学習できないことが指摘されている [***] [20]。このため、様々な情報を与えることが提案されている。例として、学習システムが所属性質問や等価性判断質問などを行なうことを認める研究が挙げられる[21]。所属性質問とは、ある文が問題となっている文法の文形式か否かを問うことであり、等価性判断質問とは、文法が問題となっている文法と同じか否かを問うことである。等価性判断質問の結果が「同一でない」という場合は、反例を提示することを要求する場合もある。これを自然言語の文法学習に応用することが考えられるが、この場合の回答者である人間は、答となる文法が未知であることや、文法観や言語観の個人差などの理由から、以上のような質問に正確に答えることはできないという問題がある。

自然言語を形式言語でモデル化して生成規則を学習することを考えると、以上に述べた困難に加えて、自然言語特有の問題が生じることが予想される。以下の3つの節では、まず自然言語学習のための正の例であるコーパスについて、次に自然言語学習における問題について、最後に我々が提案する統計的手法について述べる。

2.2 大規模言語コーパス

大規模言語コーパスの必要性は、計算言語学の分野では以前から知られており、これを作成するためのさまざまなプロジェクトがある。ここでは、本研究で

[***] 式2.1のチョムスキー階層の左に位置する言語族ほど、生成規則の制限が緩和されるため文法学習が困難になる。

用いた英語コーパス Penn Treebank [2) 3) 4)] を例にとり、付加される情報とコーパス作成について述べる。

2.2.1 付加情報

言語コーパスには、文章のみならず様々な情報が付加されていることがある。Penn Treebank には、以下の 2 つの情報が付加されている。

- 品詞情報 (part-of-speech tag information, POS tag information)
- 統語構造情報 (syntactic structure information)

品詞情報は、各単語に与えられ、それぞれの単語の品詞を表す。Penn Treebank で用いられる品詞タグ一覧を表 .1 に掲げる。統語構造情報は各文に与えられ、その文の構造を表す。以下の例は Penn Treebank に含まれるコーパスの 1 つ Air Travel Information System transcripts (ATIS) [22)] に含まれる 1 つの文とその付加情報である。

```
Does flight 975 serve dinner ?
VBZ  NN     CD  VB    NN     .
(SQ Does (NP flight 975) (VP serve (NP dinner) ) ? )
```

この例の一行目は ATIS に実際に存在する文であり、二行目と三行目が付加された情報である。二行目はそれぞれの単語の品詞を表している。三行目は文の統語構造であるが、この中の SQ や NP は統語タグ (syntactic tag) と呼ばれる記号で、単語の連続で構成される句の分類を表している。Penn Treebank で用いられる統語タグ一覧を表.2 に掲げる。

2.2.2 コーパスの作成

一般に、ある種の情報が付加されたコーパスを作成する場合、この情報付加は計算機を用いて半自動的に行なわれるが、計算機の出力が必ずしも正しいわけではないので、最終的には人手で確認する必要がある。Penn Treebank では品詞タグの割り当てに、まず PARTS [14)] を用いて自動で処理する。PARTS の品詞タグは Penn Treebank の品詞タグと少し異なるため変換を要するが、この変換も自動で行なう。コーパスの一部について精度を調べた結果、この段階での誤り率は 9.6% で、後続する人手による修正を経て誤り率は 5.4% となったと報告してい

る[3]。同様に、統語情報の付加についても、まず Hindle の決定的構文解析器[23][24]を用いて自動処理し、その出力を人手で修正する。この結果の精度については正確に報告されていないが、これは統語解析結果の評価が困難であるためと思われる。統語解析結果の評価については 4 章でふれる。

コーパスの作成における問題としては、上で述べた精度に加えて、作成にかかるコストが挙げられる。自動処理の段階にかかるコストは、人手による修正のコストに比べて極めて小さいので、人手による修正のコストがコーパス作成全体のコストとみなせる。人手による修正のコストの目安になるのが、1 単語あたりの修正時間である。Penn Treebank の作成における 1,000 単語あたりの修正時間は以下のように報告されている[3]。

- 品詞情報の修正　　22 分
- 統語構造情報の修正　　2 時間 †

この結果から、統語構造情報の修正は品詞情報の修正と比較して約 6 倍の時間を要していることがわかる。これには、品詞情報が比較的局所的に決定されるのに対し、統語構造情報が大域的な情報を必要とすることが多いためと考えられる。このため、本研究では品詞情報のみを用いて統語規則の獲得を行ない、統語構造情報を結果の評価に用いた。

2.3　自然言語の文法学習

2.3.1　自然言語の生成規則

形式言語では所属性問題に対して正しく答えることが要求されるだけであるが、形式言語を用いて自然言語をモデル化した場合、生成規則の心理学的実在性も要求される。また、一般に統語解析に続いて行なわれる意味解析に対する影響を考慮に入れる必要がある。以下の 2 つの統語構造を例としてこれを説明する。

```
( I ( have ( a pen ) ) )
( ( ( I have ) a ) pen )
```

† 統語構造情報の修正は上述したように、正解の基準により要求される水準が大きく変わる。ここに掲げたデータはその一例である。

ここで、括弧はその中の要素が一つのまとまりを構成することを意味する。最初の統語構造では“a pen”がひとまとまりとなっている。この構造から、一般的な概念を指す“pen”が“a”とともに用いられることで一つの実在する「ペン」を意味すると考えられる。これに対して、二番目の例では“I have a”がひとまとまりになり、これが“pen”と結び付くので、「何かを1つ持っている」という概念が存在し、これが“pen”と結び付くと解釈できる。このように、心理学的実在性および後続する意味処理という視点から考察すると、例に挙げた2つの統語構造の間には違いがある。自然言語の統語解析結果は後続する処理で統語構造として用いられることと、統語構造は生成規則の集合によって決定されることを考え合わせると、自然言語の生成規則は、原則的に所属性問題に答えることだけが必要な形式言語の生成規則と異なる意味を持つと結論できる。

2.3.2 文法学習

自然言語処理を行なうシステムを構成する際、一般的に形式言語を用いてモデル化することが多い。このとき文脈自由文法と正規言語が最もよく用いられる。この理由は、この言語族の性質が十分に知られていることである。十分な速度の構文解析アルゴリズムが知られているのも、式2.1のチョムスキー階層を右から見て文脈自由言語 ($\mathcal{CFL}$) までの言語族である。しかし、文脈自由文法の生成規則の正の例からの学習は、上述したように、制限を課すことなしには不可能である。自然言語の規則学習には、一般に得ることが困難な負の例 ($\overline{L(G)} = T^* - L(G)$) を用いることはできず、ノイズを含む正の例を用いることになる。また、自然言語を直観的に知っている人間に所属性質問や等価性質問を課しても、量と質の両方で十分な情報を与えることは困難である。さらに、上述したようにある2つの生成規則の集合が形式言語理論の立場から同一の言語を表すことが証明されても、自然言語処理という視点からは異なった文法となる。このように考えると、自然言語の生成規則学習に形式言語の学習のアルゴリズムを直接用いることは適切ではない。自然言語の統語規則学習に対しては、正の例である大規模な計算機可読のコーパスが近年次々と出現し、今後のこの規模や種類がますます増加することは確実であるから、大量のデータを処理する統計的手法を用いることが望ましいと考えられる。このような考えに基づいて、さまざまな言語知識を統計的に獲得する研究がなされている。以下では、これらの研究の中で統語的知識の獲得を目指した試みについて説明し、本研究の意義を明らかにする。

単語の分類

コーパス中における単語の出現状況をもとに単語を分類することができる。Fernando Pereira 他 [25] は、統語解析の結果を用いて他動詞とその目的語の主辞である名詞の関係を確率を用いて表し、名詞を分類している。この分類に際して必要となる2つの確率分布の間の類似度として Kullback-Leiber 距離 を用いている[††]。また、Resnik [26] や Hindle [27] は相互情報量を用いて同様の研究を行なっている [†††] 。本研究では品詞列の分類を確率分布の距離を用いて行なった。

動詞の格フレームの獲得

動詞の格フレームとは、ある動詞に対して与えられ、その動詞の要求する格としてどのようなものがるかを記述したものである。Brent と Berwick [28] は、英語の動詞に対してこの格フレームを5種類用意し、コーパスに出現する動詞がどの種類に属するかという知識を獲得する研究を行なった。この他に類似の研究として Brent [29] や Manning [30] の研究が挙げられる。これらの研究では、格フレームの種類をあらかじめ決めておき、構文解析において曖昧性の少ない文のみを対象にすることで高い精度を得ているが、一方で再現率が低くなることや未知の格フレームを獲得することができないという問題がある。

生成規則に対する確率付与

一般に、句構造文法の生成規則に確率値を付与したものを確率文法 (stochastic grammar) と呼ぶ。例えば、文脈自由文法に確率を付与したものは確率文脈自由文法 (stochastic context-free grammar, SCFG) と呼ばれる。Fernando Pereira と Yves Schabes [11] や Yves Schabes 他[12] は、確率文脈自由文法をチョムスキー標準

[††] 2つの確率変数 $P_1(x), P_2(x)$ に対して Kullback-Leiber 距離 $D(P_1||P_2)$ は以下のように定義される。

$$D(P_1||P_2) = \sum_x P_1(x) \log \frac{P_1(x)}{P_2(x)}$$

[†††] 2つの事象 A, B が生起する確率を $P(A), P(B)$ としたとき、相互情報量 $I(A;B)$ は以下のように定義される。

$$I(A;B) = \log \frac{P(A,B)}{P(A)P(B)}$$

形で表し、それぞれの生成規則の確率値をコーパスの統語構造情報から計算するアルゴリズムを提案し高い精度を報告している[‡]。この研究の問題点として以下のことが挙げられる。

- 獲得した規則についての考察を行なっていない。
- 新たな統語タグを得ることができない。
- 3つ以上の記号列が一つ非終端記号に書き換えられることを認めていないため、獲得した生成規則の心理学的実在性に乏しい。

変形規則の獲得

Eric Brill [13) 17)] は以下のようにして文の統語構造を解析する規則をコーパスから獲得する試みを行なった。

1. 英語が右分岐の言語であることを利用して、すべての文に右分岐の構造を与える。
2. ある部分の前後の品詞や単語の情報を利用して構造を変更する規則を複数用意する。
3. 規則の候補の中から、コーパスに適用することで、付加されている構造に最も近付ける規則を選び、これをコーパスに適用する。
4. ある条件が満たされるまで3を繰り返す。

この結果得られる変形規則[‡‡]の解釈は困難であり、発展性に乏しいことが問題であると考えられる。

[‡] 文脈自由文法の生成規則が次の 1), 2), 3) の形式のいずれかに限定されているとき、その文法はチョムスキー標準形であるという。

$$1)\ A \to B \cdot C,\ \ 2)\ D \to a,\ \ 3)\ S \to \varepsilon \quad (A, B, C, D \in N,\ a \in T)$$

任意の文脈自由文法は、それと同じ言語を生成するチョムスキー標準形の文脈自由文法に変換可能である。

[‡‡] ここでいう変形規則は、変形生成文法で用いられる変形規則とは異なる

品詞列の置き換え

Eric Brill と Mitchell Marcus [31] は、コーパスを品詞の連接と見なして、ある長さ 2 の品詞列と他の長さ 1 の品詞列との類似性を Kullback-Leiber 距離 ‡‡‡ を用いて計算し、最も類似する長さ 1 の品詞列で長さ 2 の品詞列を置き換える規則を文法として獲得するということを試みている。例えば、$DT \cdot NN$ に最も類似している品詞が PRP であることから、$PRP \rightarrow DT \cdot NN$ を生成規則として獲得する。この方法には以下のような問題がある。

- 3 つ以上の品詞列が一つ非終端記号に書き換えられることを認めていないため、獲得した生成規則の心理学的実在性に乏しい。
- 非終端記号として終端記号を用いるため、対応する終端記号がない場合や下位範疇化を必要とする場合に対応できない。

この問題点を克服するため、本論文では任意の長さの品詞列を対象として類似性を計算し、新たに導入する同一の非終端記号に書き換えることを提案する。ただし、次の節で述べるように、我々が用いた類似度の計算式はこの研究で用いられたものとは異なる。

一般化相互情報量を用いた構文解析

Eric Brill 他[32] は、複数の品詞列の間の結合の強さを一般化相互情報量を用いて表し、最も結合が弱い位置で文を2つに分割することを再帰的に繰り返すことで、文の統語的構造を解析することを提案している* 。David Magerman と Mitchell Marcus [33] はこれを Brown Corpus [34] に対して用いた結果を報告している 。一般に前置詞と冠詞の間の一般化相互情報量は高いことからこの 2 つの品詞の間の結

‡‡‡ Eric Brill と Mitchell Marcus が提案している類似度は Kullback-Leiber 距離を拡張した

$$D(P_1||P_2) + D(P_2||P_1) = \sum_x P_1(x) \log \frac{P_1(x)}{P_2(x)} + \sum_x P_2(x) \log \frac{P_2(x)}{P_1(x)}$$

である。対称律が成り立つ点で Kullback-Leiber 距離より妥当性があると考えられる。

* ある 2-gram xy の文脈 $x_1 x_2 x_{i-1} x y y_1 y_2 y_{j-1}$ における一般化相互情報量は、k_X を重み関数として

$$\mathcal{GMI} = log \frac{\prod_{X\ crosses\ xy} k_X P(X)}{\prod_{X\ dose\ not\ crosses\ xy} P(X)^{(i+j)/2}}$$

で定義される。

合は強いと判断され、ここで分割されることが少ない。これを避けるために、結合しない2つの品詞をあらかじめ与えることを行なっているが、この統語的知識の根拠が明確ではなく一般化相互情報量のみを用いる枠組に問題がある。

本研究の問題定義

本研究では、以下のように問題を設定した。

- 言語を形式言語でモデル化する。
- コーパスを品詞の連接とみなす。
- 品詞情報のみから生成規則を獲得する。

用いる情報を品詞情報に限定した理由は、コーパスの作成のコストが比較的低いことと、人間の言語直観をできるだけ排除しようと考えたことである。上述した先行研究と同様に我々も統計を用いた。統計的手法は、自然言語の本質的曖昧性やコーパスに含まれる誤りに対処できると考えられる。次の節で、我々が用いた統計的手法について説明する。

2.4 統計的手法

この節では、我々が実際に用いた言語に対する統計について説明する。まず、本論文で用いる記号を定義しておく。

本研究ではコーパスを品詞の連接と見なすので、終端記号は表.1 で与えらる記号である。終端記号の集合を T で表し、その要素を pos で表す。

$$T = \{CC, CD, \ldots, .\} \quad (\text{最後のピリオドも品詞である})$$

同様に、非終端記号の集合を N で表し、その要素を syn で表す。本研究では非終端記号は統計計算に伴って導入されるので、表 .2 に掲げた Penn Treebank の統語タグ一覧ではないことに注意しなければならない。終端記号と非終端記号を合わせて tag で表す。以上から

$$pos \in T, \quad syn \in N, \ tag \in N \cup T$$

である。また、ボールド文字はそれぞれの記号の長さ 1 以上の連接を表す。一般に $X^+ = X^* - \{\varepsilon\}$ として

$$\boldsymbol{syn} \in N^+,\ \boldsymbol{pos} \in T^+,\ \boldsymbol{tag} \in (N \cup T)^+$$

が成り立つ。本研究では文を品詞列とみなし、コーパスを文の連接とみすので、コーパスは $\boldsymbol{pos}_C = \boldsymbol{pos}_1 \cdot \boldsymbol{pos}_2 \cdots \boldsymbol{pos}_n$ と表すことができる。ここで、$\boldsymbol{pos}_i\ (i = 1, 2, ..., n)$ は文である。ある記号列 $\boldsymbol{tag}_1$ がもう 1 つの記号列 $\boldsymbol{tag}_2$ の部分文字列となっていることを $\boldsymbol{tag}_1 \in \boldsymbol{tag}_2$ で表す。例えば、コーパス中の任意の品詞列 $\boldsymbol{pos}_j$ に対して $\boldsymbol{pos}_j \in \boldsymbol{pos}_C$ が成り立つ。

2.4.1 出現頻度

ある品詞列 $\boldsymbol{pos}$ の出現頻度はコーパス中の品詞列 $\boldsymbol{pos}$ の数であり、$f(\boldsymbol{pos})$ で表す。これは $n = |\boldsymbol{pos}|$ として、連続する n 個の記号の頻度であるから、コーパスに対して n グラム統計 (n-gram statistics) を行なうことで得られる。

2.4.2 出現確率

ある品詞列 $\boldsymbol{pos}$ の出現確率を $P(\boldsymbol{pos})$ で表し、以下のように定義する。

$$P(\boldsymbol{pos}) = \frac{f(\boldsymbol{pos})}{\sum_i f(\boldsymbol{pos}_i)} \quad (|\boldsymbol{pos}_i| = |\boldsymbol{pos}|,\ \boldsymbol{pos}_i \in \boldsymbol{pos}_C) \tag{2.2}$$

ここで、分母は長さが $|\boldsymbol{pos}|$ に等しい品詞列のコーパス中の出現頻度の和である。これは $|\boldsymbol{pos}_C| - |\boldsymbol{pos}| + 1$ に等しく、さらに $|\boldsymbol{pos}_C| \gg |\boldsymbol{pos}| - 1$ であるから 式 2.2 は

$$P(\boldsymbol{pos}) = \frac{f(\boldsymbol{pos})}{|\boldsymbol{pos}_C| - |\boldsymbol{pos}| + 1} \approx \frac{f(\boldsymbol{pos})}{|\boldsymbol{pos}_C|} \tag{2.3}$$

となる。この式を用いることで出現確率が出現頻度から簡単に計算できる。

2.4.3 確率分布

事象 A, B に対して、事象 B が生起するという条件のもとでの事象 A が生起する確率を条件付き確率といい、以下の式で定義される。

$$P(A|B) = P(A \cap B)/P(B) \tag{2.4}$$

ここで、$P(B) \neq 0$ を仮定している。本研究では、ある n-gram が与えられたときの $(n+1)$-gram の出現を条件付き確率で数値化し、これを n-gram の確率分布として用いた。以下でこの確率分布の定義を述べる。

ある品詞列 $\boldsymbol{pos}$ に pos_i が先行する確率 $Pl_i(\boldsymbol{pos})$ 及び、pos_i が後続する確率 $Pr_i(\boldsymbol{pos})$ は条件付き確率であるから、以下のように表される。

$$Pl_i(\boldsymbol{pos}) = P(pos_i \cdot \boldsymbol{pos} \mid \boldsymbol{pos})$$
$$Pr_i(\boldsymbol{pos}) = P(\boldsymbol{pos} \cdot pos_i \mid \boldsymbol{pos})$$

ここで、式 2.4 および式 2.3をこの順に用いて

$$Pl_i(\boldsymbol{pos}) = \frac{P(pos_i \cdot \boldsymbol{pos})}{P(\boldsymbol{pos})} \approx \frac{f(pos_i \cdot \boldsymbol{pos})}{f(\boldsymbol{pos})} \tag{2.5}$$
$$Pr_i(\boldsymbol{pos}) = \frac{P(\boldsymbol{pos} \cdot pos_i)}{P(\boldsymbol{pos})} \approx \frac{f(\boldsymbol{pos} \cdot pos_i)}{f(\boldsymbol{pos})} \tag{2.6}$$

となる。これを用いて $\boldsymbol{pos}$ の左確率分布 $\boldsymbol{Pl}$ および、右確率分布 $\boldsymbol{Pr}$ を以下のように定義する。ここで k は品詞の種類の数である $(k = |T|)$ 。

$$\boldsymbol{Pl} = (Pl_1, Pl_2, \ldots, Pl_k)$$
$$\boldsymbol{Pr} = (Pr_1, Pr_2, \ldots, Pr_k)$$

これらの値は、式 2.5 および式 2.6 を用いることで n-gram と $(n+1)$-gram の頻度から簡単に計算できる。品詞列 “DT NN”、“DT JJ NN” 及び “MD VB” の左右の確率分布を表 .3, .4, .5 に掲げる。これらの表の中の左エントロピーと右エントロピーについては次の節で定義する。

2.4.4 エントロピー

エントロピー (entropy) は情報の量や不確かさを測る量として 1948 年にシャノンによって導入された概念であり、情報源に対して付与される。いま無記憶情報源 X を考えると、これはアルファベット $T = \{a_1, a_2, \ldots, a_n\}$ に値をとる独立かつ一定の分布をもつ確率過程とみなされる。その確率分布を $P(a_i) = P_i$ とするとき、無記憶情報源 X のエントロピーは

$$H(X) = -\sum_{i=1}^{n} P_i \, log \, P_i \tag{2.7}$$

と定義される。ここで、対数の底は 2 である (以下同様) 。エントロピーは等確率分布のとき最大となることがわかっている。

本研究では、ある品詞列に先行、又は後続する品詞の確率分布に対してエントロピーを計算し、品詞列の抽出の基準として用いた。これは、次章で述べるように品詞列の独立性を表す尺度の一つであると考えられる。

ある品詞列 $\boldsymbol{pos}$ に先行する品詞の左確率分布 $\boldsymbol{Pl}$ に対するエントロピーを左エントロピーと呼び H_l で表す。同様に後続する品詞の出現確率の右確率分布 $\boldsymbol{Pr}$ に対するエントロピーを右エントロピーと呼び H_r で表す。式2.7 から、これらは以下のように計算される。

$$H_l(\boldsymbol{pos}) = -\sum_i Pl_i(\boldsymbol{pos}) \log Pl_i(\boldsymbol{pos}) \tag{2.8}$$

$$H_r(\boldsymbol{pos}) = -\sum_i Pr_i(\boldsymbol{pos}) \log Pr_i(\boldsymbol{pos}) \tag{2.9}$$

ここで、総和は $pos_i \in T$ となるすべての i に対して計算されるものとする (以下同様)。

表 .6 は Penn Treebank に含まれる Wall Street Journal の記事の品詞列に対して左エントロピーと右エントロピーを計算し、この和 $(H_l + H_r)$ の大きい順に並べた品詞列の例である。表中の V_l および V_r は次節で述べる分散である。

2.4.5 分散

分散 (variance) は確率分布のばらつきを表す量であり、確率変数 P_i $(i = 1, 2, \ldots, n)$ に対して、平均値を $\overline{P}$ として、

$$V = \sum_{i=1}^{n} (P_i - \overline{P})^2 \tag{2.10}$$

と定義される。この式から、分散は $P_i = \overline{P}$ $(i = 1, 2, \ldots, n)$ のとき最小となることが示せる。この定義から分散 V は $\boldsymbol{P} = (P_1, P_2, \ldots, P_n)$ として

$$V = |\boldsymbol{P}|^2 - n\overline{P}^2 \tag{2.11}$$

計算できることが示せる (付録 A.1 参照) 。

本研究では、ある品詞列に先行、又は後続する品詞の確率分布に対して分散を計算した。これはエントロピーと同様に品詞列の独立性を表す尺度の 1 つであると考えられる。

ある品詞列 $\boldsymbol{pos}$ に先行する品詞の左確率分布 $\boldsymbol{Pl}$ に対する分散を左分散と呼び V_l で表す。同様に後続する品詞の出現確率の右確率分布 $\boldsymbol{Pr}$ に対する分散を右分散と呼び V_r で表す。これらは品詞の種類の数を $k = |T|$ とすると $n = k$, $\overline{P} = 1/k$ であるから、これを式 2.11 に代入して

$$\begin{aligned} V_l(\boldsymbol{pos}) &= |\boldsymbol{Pl}|^2 - \frac{1}{k} \\ V_r(\boldsymbol{pos}) &= |\boldsymbol{Pr}|^2 - \frac{1}{k} \end{aligned}$$

である。表 .7 は左分散と右分散の和 $(V_l + V_r)$ の小さい順に並べた、品詞列の例を示す。表中の H_l および H_r は前節で述べたエントロピーである。

2.4.6 分布分析

ある記号列の分布 (distribution) とは、その記号列のコーパスにおける文脈の集合である。この文脈として、前後に位置する一定の長さの記号列などが考えられる。前節で述べたように、分布を表す基準として前後に位置する 1 単語を用い、この条件付き確率を分析することて単語の分類を行なった研究があり、分布を統計的に処理する分布分析 (distribution analysis) の有効性が示されている。本研究ではこの分布分析を品詞列の分類に応用した。単語の分類と同様に、分布を表す基準として前後に位置する 1 品詞を選び、これを条件付き確率により数値化した。しかし、分布から類似度への写像は独自に選択した。我々が用いた類似度を表す関数は、2 つの品詞列を $\boldsymbol{pos}_1$, $\boldsymbol{pos}_2$ として、以下の式で与えられる。

$$d(\boldsymbol{pos}_1, \boldsymbol{pos}_2) = \sqrt{\sum_i \{Pl_i(\boldsymbol{pos}_1) - Pl_i(\boldsymbol{pos}_2)\}^2 + \sum_i \{Pr_i(\boldsymbol{pos}_1) - Pr_i(\boldsymbol{pos}_2)\}^2}$$

この式は、$\boldsymbol{P}_1 = (\boldsymbol{Pl}_1, \boldsymbol{Pr}_1)$, $\boldsymbol{P}_2 = (\boldsymbol{Pl}_2, \boldsymbol{Pr}_2)$ とすると $|\boldsymbol{P}_1 - \boldsymbol{P}_2|$ と等しくなる。よってこれは、2 つの品詞列の左右の確率分布をベクトルと見なしたときの距離である。Eric Brill と Mitchell Marcus が提案する式では $P(x) = 0$ となるときに問題が生じるが、我々が提案する式ではこの問題が生じないことを付記しておく。例として表.8 に長さ 5 以下の品詞列を品詞列 $DT \cdot NN$ との確率分布ベクトルの距離が小さい順に並べた結果を掲げる。

第 3 章 統計による統語規則の抽出

この章では、まず本研究で用いた自然言語の性質に対する仮定について述べる。次にこの仮定に基づき、前章で述べた統計的手法を組み合わせ、自然言語の生成規則をコーパスから獲得する方法を述べる。

3.1 統計による統語規則の抽出

3.1.1 生成規則と文形式の関係

形式言語においては、アルファベットの連接が文形式であるか否かは生成規則によって決定される。このことから、文形式の集合である言語は生成規則の特徴を反映していると考えられる。本研究では自然言語を形式言語を用いてモデル化しているので、コーパスを文型式の集合とみなすことができる。以上のことから、コーパスは自然言語の生成規則の特徴を反映していると考えられる。我々はこの考えに従って、コーパス中の品詞列は以下のような特徴をもつと仮定した。

- 生成規則の右辺に現れる品詞列は環境から独立して高い頻度で現れる。
- 同じ非終端記号から直接導出される品詞列は類似した環境を持つ。

次節以下では、この 2 つの仮定のそれぞれについて詳しく述べ、これらを利用して統語規則を獲得する手法を具体的に説明する。

3.1.2 品詞列の独立性

ある品詞列が独立であるとは、その品詞列がひとまとまりとなって意味をなすことをいう。例として以下の文について考える。

```
It  has not happened for the longest time .
PRP VBZ RB  VBN      IN  DT  JJS     NN   .
```

この文に含まれる品詞列の中で、品詞列 “DT JJS NN” は独立であり、品詞列 “VBN IN DT” は独立ではない。心理学的実在性を有する自然言語の生成規則の右辺に

は、独立な品詞列のみが現れると考えられる。そこで、生成規則を獲得するためには、まず独立な品詞列を抽出することが問題となる。我々はまず、独立な品詞列はコーパス中に高い頻度で現れると仮定した。この仮定を具体的に説明するために、例として文法 G に以下の規則が含まれている場合を考える。

$$syn_a \rightarrow \boldsymbol{pos}_a \quad (S \stackrel{*}{\Rightarrow} \boldsymbol{tag}_x \cdot syn_a \cdot \boldsymbol{tag}_y) \tag{3.1}$$

このとき、$\boldsymbol{pos}_a$ を部分文字列としてもつ文が言語 $L(G)$ の要素として存在する。コーパス C は言語 $L(G)$ の部分集合であるから、大数の法則から C の大きさがある程度大きくなれば $\boldsymbol{pos}_a$ を部分文字列として含む文がコーパス中に多数存在すると考えられ、$\boldsymbol{pos}_a$ の出現頻度 $f(\boldsymbol{pos}_a)$ が高いことが期待できる。よって、独立な品詞列は高い出現頻度を持つ可能性が高いと考えられる。しかし、生成規則の右辺にない品詞列も文の部分文字列となることがあるので、この出現頻度についての議論が必要である。例えば、$\boldsymbol{pos}_x = \boldsymbol{pos}_{x1} \cdot \boldsymbol{pos}_{x2}$ を G のいかなる生成規則の右辺にもあらわれない品詞列とし $(\forall syn \rightarrow \boldsymbol{pos}_x \notin P)$、

$$\begin{aligned} syn_b &\rightarrow syn_c \cdot syn_d \quad (S \stackrel{*}{\Rightarrow} \boldsymbol{tag}_x \cdot syn_b \cdot \boldsymbol{tag}_y) \\ syn_c &\rightarrow \boldsymbol{pos}_{c'} \cdot \boldsymbol{pos}_{x1} \\ syn_d &\rightarrow \boldsymbol{pos}_{x2} \cdot \boldsymbol{pos}_{d'} \end{aligned}$$

が生成規則 P の要素であるとすると、

$$syn_b \stackrel{*}{\Rightarrow} \boldsymbol{pos}_{c'} \cdot \underbrace{\boldsymbol{pos}_{x1} \cdot \boldsymbol{pos}_{x2}}_{\boldsymbol{pos}_x} \cdot \boldsymbol{pos}_{d'} \tag{3.2}$$

であることから、$\boldsymbol{pos}_x$ が文の部分文字列として出現し得ることがわかる。このとき出現頻度 $f(\boldsymbol{pos}_x)$ は、$\boldsymbol{pos}_x$ を部分文字列として含む品詞列を導出する非終端記号が制限されているため、あまり高くなることはないと考えられる。しかし、ある独立な品詞列の頻度より大きくなる可能性がある。実際 Wall Street Journal では、$f(DT \cdot JJS \cdot NN) = 257,\ f(VBN \cdot IN \cdot DT) = 7087$ であり、この逆転が生じていることがわかる。よって、頻度だけで品詞列の独立性を判断することは適切ではない。そこで我々は、品詞列の左右に現れる品詞群の多様性に注目した。式 3.1 と 3.2 を比べると $\boldsymbol{pos}_a$ の左右に現れることができる品詞群には制限がないが、$\boldsymbol{pos}_x$ の左右に現れることができる品詞は、これが syn_b から導出される場合には、それぞれ $\boldsymbol{pos}_{c'}$ の右端に位置する品詞、$\boldsymbol{pos}_{d'}$ の左端に位置する品詞

に制限される。他の非終端記号から導出される場合も同様の制限を受ける。このことから、独立な品詞列の左右に位置する品詞群は、独立でない品詞列の左右に位置する品詞群より多様であると考えられる。我々は、この多様性を表す量として、品詞列の左右の確率分布のエントロピー、又は分散を用いることを考えた。エントロピーは $\boldsymbol{pos}$ の左右の品詞の確率分布の情報量であり、分散は等確率分布との距離である。表 .6 と表 .7 を比べることで、この二つの尺度が品詞列の独立性を表す限りにおいて、大きな差異がないことが分かる。本研究では、品詞列の独立性を表す尺度として左右のエントロピーの和 $(H_l + H_r)$ を選択した。この理由は以下の通りである。

> エントロピーは大きいほど独立であると考えられ、分散は小さいほど独立であると考えらる。注目している品詞列の頻度を f とすると、それぞれの限界である最大値および最小値は等確率分布に対して与えられ、$H_{max} = \log f, \quad V_{min} = 0$ である (付録 A.2, A.3 参照)。エントロピーの最大値が頻度に対して狭義の単調増加となるのに対して、分散は頻度に対して一定である。すでに述べたように、独立性として頻度も考慮に入れるべきであるから、独立性を表す尺度として、エントロピーが分散よりも優れている。

独立性の尺度として、以上に述べたエントロピーに加えて、左右の記号の割合を提案する。ここでいう記号とは、具体的にはピリオドやコンマなどのような単語に対応しない品詞を意味する。本研究ではコーパスを文の連接として定義したので、文の先頭に位置する品詞列の左隣と、末尾に位置する品詞列の右隣は、以下の例が示すように、かならず文区切りを表す品詞 “.” である。

```
go there . Does flight 975 serve dinner ? And does it
VB PRP   . VBZ  NN     CD  VB    NN     . CC  VBZ  PRP
         ^^^^^^^^^^                ^^^^^^^^
```

また、独立な品詞列がコンマなどの記号 (以下、デリミターと呼ぶ) を含むことは少なく、文が独立な品詞列とデリミターの連接として表されると考えられるので、これらの記号 に隣接する品詞列は独立である可能性が高い。よって、品詞列の独立性の基準として左右のデリミターの割合を利用することができる。本研究の実験では、文末記号とコンマ以外の記号を含む文を対象としなかったので、デリミターは “,” と “.” である。

以上、2つの独立性の基準について説明した。品詞列の抽出を実際に行なう際は、これらの値に閾値を設け、閾値との大小関係によって独立であるか否かを決定した。また、一定の出現頻度がないと統計値が信頼できないので出現頻度にも制限を設けた。以上をまとめると、現頻度の閾値を f_{min}、エントロピーの閾値を H_{min}、デリミターの条件付き確率の閾値を Pd_{min} として、品詞列 $\boldsymbol{pos}$ が独立である条件は以下の不等式をすべて満たすことである。

1. $f(\boldsymbol{pos}) \geq f_{min}$
2. $H_l(\boldsymbol{pos}) \geq H_{min}$
3. $H_r(\boldsymbol{pos}) \geq H_{min}$
4. $Pd_l(\boldsymbol{pos}) = P(\,''.'' \cdot \boldsymbol{pos} \mid \boldsymbol{pos}) + P(\,'','' \cdot \boldsymbol{pos} \mid \boldsymbol{pos}) \geq Pd_{min}$
5. $Pd_r(\boldsymbol{pos}) = P(\boldsymbol{pos} \cdot\, ''.'' \mid \boldsymbol{pos}) + P(\boldsymbol{pos} \cdot\, '','' \mid \boldsymbol{pos}) \geq Pd_{min}$

例えば、$f_{min} = 0$, $H_{min} = 3$, $Pd_{min} = 0.05$ とした場合、以上の条件を満たす品詞列は 429 個であった。これらの中で頻度が高い順に 20 個を表 .1に掲げる。

3.1.3 品詞列の類似性

一般に、品詞列の中には類似する文法役割を果たすものがあると考えられており、文脈自由文法をモデルとした場合は、これを認めた上で生成規則の獲得を行なった。生成規則の獲得を行なう際に、同じ非終端記号から直接導出される品詞列は類似した環境を持つことを仮定した。例として文法の生成規則 P に以下の規則が含まれている場合を考える。

$$syn_e \rightarrow \boldsymbol{pos}_{e1}, \quad syn_e \rightarrow \boldsymbol{pos}_{e2} \quad (S \stackrel{*}{\Rightarrow} \boldsymbol{tag}_x \cdot syn_e \cdot \boldsymbol{tag}_y)$$

これは、2つの品詞列 $\boldsymbol{pos}_{e1}$ と $\boldsymbol{pos}_{e2}$ の左右に位置することができる品詞列はどちらの品詞列の場合も syn_e の左右に導出することができる品詞列であることを意味する。そこで大数の法則を用いると、$\boldsymbol{pos}_{e1}$ と $\boldsymbol{pos}_{e2}$ の出現頻度がある程度大きければ、それぞれの左右の品詞列の確率分布は互いに類似するものと考えられる。これを逆に考えることで以下の推論が成り立つ。

> $\boldsymbol{pos}_{e1}$ と $\boldsymbol{pos}_{e2}$ の出現頻度はコーパスの増大にともなって大きくなるので、大規模なコーパスに対して統計を行なった結果、これらの現れる環

境に対してある程度の類似性が認められれば、$\boldsymbol{pos}_{e1}$ と $\boldsymbol{pos}_{e2}$ を導出する同一の非終端記号 syn_e があり、 $syn_e \rightarrow \boldsymbol{pos}_{e1}$, $syn_e \rightarrow \boldsymbol{pos}_{e2}$ が生成規則に含まれる。

本研究では、以上の仮定を用いて品詞列の左右の 1 品詞の確率分布をベクトルとみなし、2 つの品詞列の類似度がこの距離で表されるとして、複数の品詞列に対してクラスタリングを行なった。このクラスタリングは以下の 4 つ処理で構成される。

1. 独立な品詞列を抽出し、それぞれの類似度を計算する。
2. ノードに品詞列が、アークに品詞列の類似度が対応するグラフを生成する (完全無向グラフ)。
3. ある閾値以下の値を持つアークを消去する。
4. グラフの連結性を調べ、連結成分に分割する。

この結果分割された部分グラフがクラスターに対応し、それぞれのノードが要素に対応する。図.1 は、これらの各ステップを概念的に表している。例として、独立性の条件を $H_{min} = 3$, $Pd_{min} = 0.05$ とし、類似度の閾値として $D_{min} = 0.25$ を用いた場合の結果を表 .2 に掲げる。表中の「出現合計」は、各クラスターに与えられる数値で、要素である各品詞列のコーパス中の出現頻度と長さの積の合計である。これを計算する際に、同一クラスターに属する要素の中に一方が他方の部分文字列になる場合があるが、ある品詞列の出現頻度計算する際には同一クラスター内にその部分文字列が含まれる場合はこの頻度を引いて考えることとした。よって、出現合計はクラスターがコーパス中に占める記号数と見なすことができ、各クラスター間の出現合計の大小関係は、それぞれから得られる生成規則をコーパスに適応した場合のコーパスの長さ $|\boldsymbol{pos}_C|$ の減少量の大小関係を表すと考えられる * 。

以上のようにして得られる品詞列のクラスターは、同一の非終端記号に書き換えられると考えられる。つまり、ある非終端記号を導入し、左辺にこの非終端記

* 各クラスター間の出現合計の大小関係とコーパスの長さの減少量の大小関係は、必ずしも同値ではない。これは、ある同一の品詞列をそれぞれ左端と右端にもつ品詞列が含まれる場合 (例 : "DT NN" と "NN NNS") などがあるためである。しかし、大小関係に逆転が起こる場合はまれであり、これが起こったとしても以後の処理に対する影響は少ないと考える。

号を、右辺に得られた品詞列を持つ生成規則が文法 G の生成規則に含まれると考えられる。これを具体的に説明するために例として表.2 の一番上にあるクラスターをとる。非終端記号として syn_1 を導入すると、このクラスターから以下の生成規則群が得られる。

$$\begin{aligned} syn_1 &\rightarrow VBD \\ syn_1 &\rightarrow RB \cdot VBD \end{aligned}$$

この例では出現合計が 20,260 である。これは、この規則をコーパスに適用すると最大 20,260 個の記号が書き換えの対象となることを意味する。

3.2 抽出と適用の反復

前節で述べた方法で、各クラスターから生成規則群を得ることができる。しかし、品詞列であるコーパスに対してこの方法を適用して得られる生成規則は、$syn \rightarrow \boldsymbol{pos}$ のような形の生成規則に限られる。つまり、コーパスは終端記号の連接であることと、ある記号列に対して上述の統計を用いることで抽出される記号列は対象となった記号列の部分文字列となることから、コーパスに対する統計からは、右辺が終端記号の連接である生成規則しか得ることができない。しかし、自然言語の生成規則には、右辺に非終端記号を持つものもあり得る。これらの規則を獲得するために、生成規則を適用した結果得られる非終端記号を含む記号列もコーパスとみなして、統計・獲得・適用の処理を繰り返し行なうことを考えた。この繰り返しは以下の処理で構成される (図 .2 参照) 。

1. コーパスに対して n-gram 統計をとる。
2. n-gram 統計の結果をもとに生成規則を獲得する。
3. 得られた生成規則をコーパスに適用し、コーパスの状態を変化させる。
4. 処理 1 へ戻る。

以下では、これらの 3 つの処理を詳しく説明する。

3.2.1 n-gram 統計の計算

n-gram 統計には Nagao と Mori が提案したアルゴリズムを用いた。このアルゴリズムでは、まずコーパスをソートし、その結果を順に見ていくことで n-gram の頻度を効率良く計算している。ある n-gram の左右のエントロピーは、その n-gram の出現頻度と $(n+1)$-gram の出現頻度を比べることで計算されるが、我々は Nagao と Mori が提案したアルゴリズムを拡張し、頻度と左右のエントロピーを同時に計算することを可能とした。この拡張によって、エントロピーを求めるための n-gram の出現頻度と $(n+1)$-gram の出現頻度の比較にかかる計算コストを削減できた。この段階での出力はコーパス中に現れるすべての n-gram の頻度と左右のエントロピーである。

3.2.2 生成規則の抽出

本研究では、以下の 2 つの方法を用いて生成規則の獲得の実験を行なった。

- **方法** 1 ある条件を用いて抽出された 1 つの記号列を右辺とし、新たに導入した非終端記号を左辺とする。
- **方法** 2 分布に類似性がある複数の品詞列を抽出しこれを右辺とする。左辺の非終端記号は以下のように決定される。
  - 右辺の品詞列群の中に 1 つの非終端記号からなる品詞列が 1 つ含まれるならば、これを左辺の非終端記号とし右辺の品詞列群から取り除く。
  - 上記の条件が満たされない場合は、新たに非終端記号を導入し左辺とする。

方法 1 を用いて得られる生成規則は再帰することがないので、この結果得られる文法は正規文法となる。方法 2 を用いると再帰的な生成規則が得られる可能性があるが、生成規則に文脈依存性はないのでこの結果得られる文法は文脈自由文法となる。以下では、それぞれの方法における品詞列の抽出について詳しく述べる。

方法 1

この方法では、1 つの記号列を抽出する基準が問題になる。本研究では、条件

- 頻度がある閾値 (f_{min}) 以上である
- 左右のエントロピーがある閾値以上である

を満たす長さ 2 以上 10 以下の記号列 $\boldsymbol{tag}$ 中からコーパスに占める記号数 $f(\boldsymbol{tag}) \cdot |\boldsymbol{tag}|$ が最大となるものを選んだ。これは、コーパスを最も効率良く縮小する生成規則が心理学的実在性の高い文法を構成すると考えたことによる。

方法 2

この方法では複数の品詞列を同一の非終端記号に書き換える生成規則を獲得することが目的であるから、まずある条件を満たす 1 以上 10 以下品詞列を抽出しクラスタリングによって品詞列を分類する。この結果として得られるすべてのクラスタを右辺とする生成規則を生成規則の元とすることが考えられるが、それぞれの生成規則群が互いに関係することがある。例えば、表.2 の中の 2 番目のクラスターの最初の要素から、

$$syn_2 \rightarrow DT \cdot NN \tag{3.3}$$

が得られ、3 番目のクラスターの 3 番目の要素から、

$$syn_3 \rightarrow IN \cdot DT \cdot NN \tag{3.4}$$

が得られる。ここで、syn_2 と syn_3 は、2 番目のクラスターと 3 番目のクラスターのために導入される非終端記号である。この例を見ると、生成規則 3.4 は、生成規則 3.3 を用いて、

$$syn_3 \rightarrow IN \cdot syn_2$$

とすることができる。この例が示すように、それぞれの生成規則群が互いに関係することがあるので、抽出されたすべてのクラスターを生成規則に変換することは適切な生成規則の獲得を妨げることになると考えられる。この理由から、唯一のクラスターを選択する方法を選択した。次に問題となるのがクラスターの選択基準であるが、本研究では出現合計が最大となるクラスターから生成規則を獲得することとした。出現合計が大きいということは、そのクラスターが大域的構造に関係なく現れる基本的な記号列であることを意味する。クラスタリングの結果の例として掲げた表 .2 の計算に用いた条件で出現合計が最大となったのは 2 番目のクラスターであり、従来の文法における名詞句がこれに相当すると考えらる。

従来の文法における名詞句は文の基本的な構成要素であり文の大域的構造との関係は少ない。これは上記の推論を傍証する結果である。

後続する処理でこのようにして得られた生成規則をコーパスに適用するが、これを行なうことでコーパスに変化が生じる。この結果、クラスタリングで選択されなかった記号列の確率分布が変化し、記号列の間の類似性に変化が生じる可能性がある。例として、選択されたクラスターに記号列 $DT \cdot NN \cdot NN$ が含まれ、記号列 $DT \cdot NN$ が含まれず、選択されたクラスターに属する記号列の特徴の一つとしてその右に品詞 NN が出現する確率 $P(\boldsymbol{pos} \cdot NN \mid \boldsymbol{pos})$ が低いことが挙げられる場合を考える。クラスタリングの段階での品詞列 $DT \cdot NN$ の右確率分布の計算では、記号列 $DT \cdot NN \cdot NN$ がひとまとまりとなることが分かっていなかったので、$P(DT \cdot NN \cdot NN \mid DT \cdot NN)$ の値は高くなるが、クラスター抽出の結果 $DT \cdot NN \cdot NN$ がひとまとまりの記号列として書き換えられることが分かったので、記号列 $DT \cdot NN$ の右確率分布の計算においてこれを考慮するべきである。コーパスの中のすべて箇所で記号列 $DT \cdot NN \cdot NN$ が書き換えの対象となるかはこの段階で全く分からないので、これを除いた場合の $P(DT \cdot NN \cdot NN \mid DT \cdot NN)$ の値を計算することはできない。そこで、コーパスの長さの減少量が最大になる生成規則が適切であるという仮定を用いて、コーパスの中のすべて箇所で記号列 $DT \cdot NN \cdot NN$ がある非終端記号に書き換えられるとし、$P(DT \cdot NN \cdot NN \mid DT \cdot NN) = 0$ とした。具体的には、クラスタリングによって得られた生成規則をコーパスに暫定的に適用し、その結果得られるコーパスに対して再度確率分布を計算する手法を用いた。このとき候補となる記号列の条件は、独立な記号列に類似する記号列は独立であると考えられるので、統計結果の信頼性が限界となるある出現頻度 F 以上であることのみとした。この結果得られる記号列の確率分布に対して生成規則の獲得の際に導入された非終端記号の確率分布との距離を計算し、これが閾値 D_{min} 以下となる記号列を新たにクラスターに加えることとした。このとき、この記号列が非終端記号を含む場合などがあり得るので、この結果得られる記号列を以下のように分類し、それぞれに対して異なった処理を行なった。ここで、クラスタリングによって獲得した生成規則を $syn_c \rightarrow \boldsymbol{tag}_i \ (i = 1, 2, \ldots, k)$ 、これをコーパスに適用した結果、 syn_c との類似性が条件を満たす記号列を $\boldsymbol{tag} = tag_1 \cdot tag_2 \cdots tag_n$ とする。

1. $syn_c \notin \boldsymbol{tag}$ (獲得した生成規則の左辺の非終端記号を含まない場合)

処理：得られた記号列を右辺とする生成規則 $syn_c \rightarrow \boldsymbol{tag}$ を獲得する。

2. $\boldsymbol{tag} = syn_c \cdot tag_1$ (獲得した生成規則の左辺の非終端記号を左端に含み、長さが2である場合)

   処理：syn_c をクラスタリングによって得られた生成規則の右辺で置き換える。この結果得られる記号列の中でコーパス中に存在する記号列を右辺とする生成規則 $syn_c \rightarrow \boldsymbol{tag}_i \cdot tag_2$ $(\boldsymbol{tag}_i \cdot tag_2 \in \boldsymbol{tag}_C,\ i = 1, 2, \ldots, k)$ を獲得する。

3. $\boldsymbol{tag} = tag_2 \cdot syn_c$ (獲得した生成規則の左辺の非終端記号を右端に含み、長さが2である場合)

   処理：syn_c をクラスタリングによって得られた生成規則の右辺で置き換える。この結果得られる記号列の中でコーパス中に存在する記号列を右辺とする生成規則 $syn_c \rightarrow tag_1 \cdot \boldsymbol{tag}_i$ $(tag_2 \cdot \boldsymbol{tag}_i \in \boldsymbol{tag}_C,\ i = 1, 2, \ldots, k)$ を獲得する。

4. 上記以外の場合。

   処理：いかなる生成規則も獲得しない。

以上のようにして更新される生成規則群を再びコーパスに適用し、新たな生成規則が得られなくなるまでこれを繰り返す。最後に獲得された生成規則群の右辺に含まれる単独の非終端記号の数を数え、これが1であった場合、生成規則群の左辺にある暫定的に導入された非終端記号をこの単独の非終端記号で置き換え、この結果生じる左辺と右辺が等しい生成規則を取り除く。以上処理の結果、同一の非終端記号に書き換えられる記号列に対する生成規則が獲得される。

3.2.3 コーパスへの生成規則の適用

コーパスへの生成規則の適用とは、コーパスの中で生成規則の右辺にある記号列と一致する部分を同一の生成規則の左辺にある記号列で置き換えることである。このとき、複数の生成規則が適用できる場合や、同一の生成規則が部分的に重複する2箇所に対して適用可能な場合がある。このような場合は可能な書き換えの中から唯一の書き換えが排他的に選択されるとした。以下では方法1および方法2によって得られる生成規則の適用順序について述べる。

方法 1

この方法で得られる生成規則は唯一であるから、複数の生成規則の競合は起こり得ないが、生成規則が部分的に重複する 2 箇所に対して適用可能な場合は起こり得る。例えば、生成規則が $syn_1 \rightarrow tag_1 \cdot tag_1$ であり、コーパスに以下の記号列が存在すると仮定すると、

$$\ldots tag_1 \cdot tag_1 \cdot tag_1 \ldots$$

どちらの $tag_1 \cdot tag_1$ を syn_1 に書き換えるかで競合する。このような場合も含めて、本研究では単純に左から書き換えを行なうことにした。この例では、

$$\ldots syn_1 \cdot tag_1 \ldots$$

となる。このような書き換えにより、獲得される文法は左分岐の傾向をもつが、実験の結果得られた文法は右分岐の傾向を持つもので、我々の英語に対する言語直観と一致する。このことは、書き換えを左から行なうか右から行なうかは我々が提案する文法獲得手法に本質的な影響を与えなかったことを意味する。

方法 2

この方法では、同一の生成規則の競合に加えて、複数の生成規則が競合する場合が起こり得る。同一の生成規則の競合は、方法 1 と同じように左から書き換えを行なうことで解消することとした。複数の生成規則が競合する場合には、右辺がより長い生成規則を優先して適用することとした。これは、生成規則群の選択と同様にコーパスの長さの減少量が最大になる生成規則が適切であるという仮定による。例として、2 つの生成規則

$$syn_1 \rightarrow tag_1 \cdot tag_2 \tag{3.5}$$

$$syn_1 \rightarrow tag_1 \cdot tag_2 \cdot tag_2 \tag{3.6}$$

を仮定し、コーパスに以下の記号列が存在すると仮定すると、

$$\ldots tag_1 \cdot tag_2 \cdot tag_2 \ldots$$

生成規則 3.5 及び 3.6 は、両方とも適用可能であるが一方を適用すると他方が適用できなくなるので、どちらか一方の生成規則を選択しなければならない。このよ

うな場合は、右辺がより長い生成規則 3.6 を優先して適用することとしたので、この例では

$$\ldots syn_1 \ldots$$

となる。

3.2.4 反復による新たな規則の獲得

得られた生成規則群を適用することでコーパスの状態が変化する。変化したコーパスに対して再び統計をとることで新たな生成規則群が得られる。しかし、この繰り返しはコーパスに含まれる記号数を減らすので、統計結果の信頼性は徐々に減少することに注意しなければならない。つまり、出現頻度が非常に低い記号列の左右のエントロピーや左右の確率分布を用いて独立性や類似性を議論することの危険性は、出現頻度の低下にともなって増大する。この点を考慮して、生成規則が新たに獲得できない場合の処理を以下のように設定した。ここで、f_{min} は独立な記号列を抽出するときの頻度の閾値であり、F は統計結果の信頼性が限界となる出現頻度で、あらかじめ与えられる。

方法 1

独立な品詞列が新たに抽出できなければ終了する。

方法 2

生成規則が新たに獲得できなければ f_{min} を 0.8 倍する。この結果

1. $f_{min} \geq F$ であれば、生成規則の獲得を実行する。
2. $f_{min} < F$ であれば、終了する。

あらかじめ設定されるパラメーターを表 .3 にまとめた。

第 4 章 実験結果と考察

この章では、前章で述べた方法を用いて実際に生成規則の獲得を行なった結果を提示し、この結果に対する評価と考察を行なう。

4.1 コーパスのデータ

実験に用いたコーパスは Penn Treebank に含まれる Wall Street Journal である。これは、実在する雑誌の文章を計算機可読にしたコーパスであり、括弧やその他の記号を含む文もある。人間がこれらの記号を含む文章を読むときは記号の対称性などの視覚的情報も用いると考えられるが、本研究ではこれを対象外としているので、実験を行なう前に文末記号とコンマ以外の記号を含む文を取り除いた。よってコーパス中に含まれるのは、単語を表す品詞タグと文末記号とコンマを表す品詞タグである (表 .1 参照)。以上の処理の結果、コーパスに含まれる文の数は 24,678 であり、品詞の数は 549,247 であった。1 文あたりの品詞数は 22.3 である。図 .1 は文の長さの度数分布である。

4.2 実験結果と評価

実験の結果得られる出力は以下の 2 つである。

- コーパスの最終状態
- 生成規則

この 2 つを評価する方法として、人間の言語直観による文法や統語構造と比較することが考えられる。本研究の目的は現象として実在するコーパスの抽象的構造を人間の言語直観に頼ることなく分析することであるから、得られた結果が人間の言語直観と異なるものであっても、誤りと断言することはできない。また、人間の手による文法や統語構造も、文法観や言語観に個人差があることや言語自体が曖昧であることなどの理由から、確実なものとはいえない。よって、人間の手による文法や統語構造と比較することで正確な評価が可能であるとは考えられな

い。しかし、現段階で行なえることは人間の言語直観との比較のみであることも事実である。本研究では、書き換えられたコーパスを解析結果と見なしコーパスにあらかじめ付加されている統語構造と比較し、生成規則を一般的な英文法と対比することで評価した。以下では、まず評価基準を説明し、次に我々が提案する統計的手法の妥当性と上記の出力に対する評価を述べる。

4.2.1 評価基準

獲得された生成規則は順にコーパスに適用されるので、コーパスの最終状態は統語解析の結果と見ることができる。よって、2つの統語構造を比較することになる。一方は本実験の出力であり、他方は Penn Treebank に付加されている統語構造である。この比較を数値化するにあたって Black 他が提案した交差括弧数を用いた[35]。これを以下で説明する。

文脈自由文法や正規文法の統語構造は既に述べたように括弧を用いて表すことができる。このとき、括弧の位置をその括弧と文頭の間にある終端記号の数で表すと、対応する括弧対は左右の括弧の位置を表す数字の対とみなすことができる。これを、以下の例を用いて具体的に説明する。

```
( PRP ( VB ( DT NN ) ) )
0     1    2   3   4
```

一番左にある左括弧の位置は 0 であり、これに対応する右括弧の位置は 4 であるから、この括弧対は $(0, 4)$ と表せる。さらに、文に与えられた構造はこの括弧対の集合で表すことができる。これを B_1 とすると、この文の構造は以下のようになる。

$$B_1 = \{(0,4),(1,4),(2,4)\}$$

交差括弧数は、2つの括弧の集合に定義される。これを説明するために、同一の文に対して次のような異なる構造を仮定する。

```
( ( ( PRP VB ) DT ) NN )
0         1   2    3    4
```

この構造に対する括弧対の集合 B_2 は以下のようになる。

$$B_2 = \{(0,4),(0,3),(0,2)\}$$

2 つの括弧 $\boldsymbol{x}, \boldsymbol{y}$ が交差する条件は、 $\boldsymbol{x} = (x_l, x_r)$, $\boldsymbol{y} = (y_l, y_r)$ として

$$x_l < y_l < x_r < y_r \quad \text{または} \quad y_l < x_l < y_r < x_r$$

を満たすことである。上の例では、交差する括弧対の例として $(1,4) \in B_1$ と $(0,3) \in B_2$ が、交差しない括弧対の例として $(0,4) \in B_1$ と $(0,2) \in B_2$ が挙げられる。括弧対の集合 B_1のB_2 に対する交差括弧数を

$$B_1 \ni \boldsymbol{x}\ cross\ \exists \boldsymbol{y} \in B_2$$

となる $\boldsymbol{x}$ の総数と定義する。ここで $\boldsymbol{x}\ cross\ \boldsymbol{y}$ は括弧対 $\boldsymbol{x}$ と括弧対 $\boldsymbol{y}$ が交差することを意味する。

実験の結果の評価には、以下の手順で求められる値を用いた。

1. 実験によって得られた構造を B_1 、コーパスに付加された構造を B_2 として、各文毎に B_1 の B_2 に対する交差括弧数を計算する。

2. 1 の値をコーパス全体で合計し、付加した括弧の総数 (B_1 の要素数) で割る。

以下ではこの値を交差括弧数比と呼ぶ。

4.2.2 実験の条件

同じコーパスに対して条件を変えて 3 種類の実験を行なった。それぞれのパラメーターの値は以下の通りである。

- **実験** 1 方法 1 (正規文法)
  $f_{min} = 50,\ H_{min} = 3.0,\ Pd_{min} = 0.05$

- **実験** 2 方法 2 (文脈自由文法)
  $f_{min} = 2000,\ H_{min} = 3.0,\ Pd_{min} = 0.05,\ D_{max} = 0.25,\ F = 50$

- **実験** 3 方法 2 (文脈自由文法)
  $f_{min} = 2000,\ H_{min} = 3.0,\ Pd_{min} = 0.05,\ D_{max} = 0.20,\ F = 50$

それぞれのパラメーターの値は経験的に決定した。対象とする記号列の長さを制限しないことが理想的であるが、上限を大きくすると膨大な計算時間を必要とするので、これを 10 とした。結果として得られた生成規則に右辺の記号列の長さ

が 9 を越えるものが含まれていないので、この影響は特にないと考えられる。また、方法 1 では $|\boldsymbol{tag}| = 1$ である場合は無意味な記号の置き換えであるから、この可能性を排除した。方法 2 において $f_{min} = 2000$ とした理由は、これより大きくしても 2,000 前後の値になるまで複数の記号列からなるクラスターが得られなかったことである。

4.2.3 仮定を実現する手法の妥当性

本論文で提案する方法の基礎となる自然言語に対する 3 つの仮定は、それぞれ統計的手法を用いることで利用している。これらの実現方法の妥当性について以下で考察する。

品詞列の独立性

記号列が独立である必要条件として、記号列の左右の確率分布に対するエントロピーの値とデリミターの割合を用いた。表 .1 は、これらの閾値を実際に実験で用いた値に設定し、条件を満たす品詞列を抽出した結果である。表中には、我々の言語直観を基準にした場合、独立でない品詞列 (例 $NN \cdot IN$) は全く含まれていない。これは、我々が提案する品詞列の独立性に対する条件が妥当であることを示す。

品詞列の類似性

記号列の類似性を表す尺度として、記号列の左右の確率分布の距離を計算し閾値と比較することを提案した。表 .2 は、この閾値を実際に実験で用いた値に設定して、品詞列を分類した結果である。表中の各クラスターを我々の言語直観と比較すると、これらは上から順に、「動詞部分」、「名詞句」、「前置詞句」、「主語 + 動詞」に対応する。これは、左右 1 品詞に対する確率分布の距離を用いて品詞列を文法役割に応じて分類することが可能であることを示す。

生成規則の適用順序

書き換えに際しては、「コーパスの長さの減少量が最大になる生成規則が適切である」という仮定を用いたが、この妥当性を検証するために、以下の実験を行なった。

方法2の最初のループで得られる生成規則群を、仮定から導かれる順序にしたがって生成規則を適用した結果と、この優先規則を逆にして生成規則を適用した結果をそれぞれ、付加された構造と比較する。

この結果、仮定を用いた場合の交差括弧数比は 0.7% であったのに対して、仮定を逆に用いた場合の交差括弧数比は 1.1% であった。方法3に対しても同じような結果となった。これは、上述の仮定が妥当であることを示す。

4.2.4 解析結果の評価

4.2.2 項で述べた3つの条件での実験結果を表 .1 に掲げる。表中の「右線形」とは、文末記号を除いた記号列に対して最も右に位置する2つの記号を1つの記号で書き換えることを繰り返して得られる構造である。同様に「左線形」とは、これを左から順に行なうことで得られる構造である (図 .2 参照)。また、それぞれの方法による解析結果の例を付録 B に掲げる。

実験1では一つの独立な記号列を書き換えるためにかならず非終端記号を導入するので生成規則数と非終端記号数は等しくなる。文脈自由文法を仮定したほうが正規文法を仮定した場合より規則数は多いが、非終端記号数は少ないという結果になっている。これは、方法2におけるクラスタリングの後の処理で独立性を直接考慮せず、類似の記号列を獲得する手法を用いたためであると考えられ、この手法がコーパスにおける出現の偏りを補う効果を持つことを意味する。しかし、方法2における交差括弧数比が方法3のそれより大きくなるのも同様の理由によると考えられる。つまり、方法3によって獲得された文法が方法2の文法より、右辺の記号列の独立性を直接考慮せずに獲得された生成規則を多く含むためである。このことから、類似度の閾値を過度に大きすると、生成規則の精度が落ちると結論できる。

精度が最も良い方法2の交差括弧数比は、2章で述べた、コーパスの統語構造を用いた文法推論[11) 12)] の結果 (10% 前後) より少し高い。しかし、用いる情報の違いを考慮すると、この比較は必ずしも適切ではなく、逆に統語構造を用いない文法獲得が可能であることを意味すると考えられる。

4.2.5 生成規則の評価

本研究の目的は生成規則の獲得であるが、生成規則を定量的に評価することは困難である。従って、以下では一般的な英文法との比較を定性的に行なう。

それぞれの実験によって得られた生成規則の例を、表 .2, .3, .4 に掲げる。正規言語をモデルとした方法1では、多くの生成規則は妥当であると考えられるが、複数の生成規則を比較すると、不必要に非終端記号を導入していると考えられる。これは、同一の非終端記号を用いることを認めていないことが原因である。この例として、以下の生成規則が挙げられる。

$$NT015 \rightarrow IN \cdot NT003, \quad NT132 \rightarrow VBZ \cdot NT003$$
$$NT017 \rightarrow IN \cdot NT004, \quad NT241 \rightarrow VBZ \cdot NT004$$

このように、$NT003$ と $NT004$ に類似性が認められる場合は、これらを含む他の生成規則を考慮した上で、これらを同一の非終端記号とすることができる。この問題は、文脈自由文法をモデルとした方法2および方法3では解決されている。方法2および方法3の結果はかなり類似しているが、方法3によって導入された非終端記号の数は方法2によって導入された非終端記号の数より多い。これは、方法3の類似度の閾値が方法2のそれより低いため、記号列の分類がより細かくなることが原因であると考えられる。

方法3の結果である表.4 には、生成規則

$$NT002 \rightarrow IN \cdot JJR \cdot NNS \cdot NNS$$

が含まれいるが、この代りに以下の生成規則を文法に加える方がより抽象性が高いと考えられる。

$$NT001 \rightarrow JJR \cdot NNS \cdot NNS$$

このように、獲得した文法の中には再考を要する生成規則も含まれるが、導入された非終端記号の多くは我々の言語直観と対照して解釈することが可能である。表 .5 に方法2によって獲得された文法の中から文を構成する比較的単純な生成規則を掲げる。表.6 は、これらの生成規則に現れる非終端記号に対する解釈である。表.5 の中の左辺が $NT006$ である生成規則は、右辺に名詞句や前置詞句をさまざまな組合せで含む。これらは、動詞の格フレームを表すと考えられ、2章で述べた Brent と Berwick [28] や Brent [29] や Manning [30] の研究において仮定されて

いた動詞の格フレームの種類の自動獲得が可能であることを示す結果である。また、左辺が $NT009$ である生成規則は右辺に名詞句と文をこの順に含む。これは、文に修飾された名詞や、同格の文をもつ名詞句であると考えられる。これは、文の中に文が含まれるという再帰的構造があることを示している。

4.3 考察

実験の結果得られた生成規則の右辺は、いずれの方法でも、主に左に終端記号があり右に非終端記号があるという構造をしている。これは右分岐の言語の生成規則の特徴であり、英語が右分岐の言語であるという一般的な見解と一致する。この右分岐の傾向は、文脈自由文法をモデルとする方法2および方法3でより顕著である。さらに、得られた文法は単に右分岐であるだけでなく、我々の言語直観にかなり近い。しかしながら、生成規則の中には他の生成規則と比べることでより明解に記述できるものもある。また、以下のような解釈が難しいものも含まれる。

$$NT002 \rightarrow NT002 \cdot POS \cdot NT003 \cdot NT004 \quad (\text{方法2})$$

この他の問題として、品詞列の多義性の問題が挙げられる。例えば、品詞列 $VBG \cdot NN \cdot NNS$ は、一般的に文脈に応じて、名詞句、形容詞句および副詞句のいずれかでありえるが、本研究の方法ではこれを1つの非終端記号に書き換えるので、この点が明確になりにくい。また、統計的手法では出現頻度の低い生成規則を獲得することも困難である。これらの問題は、一度獲得した生成規則を総合的に見直す機構や開始記号に帰着する生成規則群を用いたトップダウン的推論[36) 37)]との融合で解決できると考える。

また、前置詞句や関係代名詞節などの修飾先は、必ずしも一箇所ではなく、同一の品詞列に対して異なる構造がありうる。このような統語的多義性の問題に対しても、統計的手法による解決が提案されている。例えば、Hindle と Rooth[38)]は、前置詞句が修飾する名詞の決定について、Cardie[39)]は関係代名詞が修飾する名詞の決定について解決法を提案している。これらは、品詞レベルでは解決できない問題であり、個々の単語を考慮に入れる必要がある。また、品詞体系も人間の言語直観であるから今後は単語や形態素を単位とした、より客観性の高い統計による文法獲得がなされるべきであるが、この場合も本論文で提案する手法が有効であると考えられる。

第 5 章 結論

本論文では品詞情報が付加されたコーパスから統計的手法を用いて自然言語の統語規則を獲得する方法について述べた。我々が提案する統語規則獲得の手法の特徴は以下の 3 点である。

1. 比較的単純な形式言語を用いて自然言語をモデル化した。

2. 付加するコストが低い品詞情報のみを用いる。

3. 統計的手法によるためコーパスに含まれる誤りの影響を受けにくい。

上述の点は人間の言語直観の影響を排除した文法をコーパスから自動的に獲得するために有益であると考えられ、本研究の意義はこの点にある。

本研究では、自然言語のモデルとして形式言語を用いているが、自然言語の生成規則の心理学的実在性やコーパスに含まれる誤りなどを考慮すると、形式言語の学習理論を直接用いることはできない。本論文では自然言語の文法獲得を行なう際に有効となる以下の 3 つの仮定を提案し、この仮定に基づいて生成規則獲得を行なった。

1. 生成規則の右辺に現れる品詞列は環境から独立して高い頻度で現れる。

2. 同じ非終端記号から直接導出される品詞列は類似した環境を持つ。

3. コーパスの長さの減少量が最大になる生成規則が適切である。

1 の仮定は、品詞列の出現頻度と左右のエントロピーを用いることで実現し、生成規則の候補の抽出に用いた。2 の仮定は、複数の品詞列の左右の品詞の確率分布を類似度に写像することで実現した。これら 2 つの仮定は生成規則の候補の抽出に用いた。3 の仮定は、獲得された生成規則群の選択と適用における競合を解消するために用いた。

獲得された文法には、我々の言語直観に適合する生成規則が多数含まれていた。人間の言語直観には客観的な根拠がないため、このことが必ずしも獲得された生成規則の妥当性を証明するものではない。しかしながら、比較的単純なモデ

ルと上述の3つの仮定から右分岐の傾向を持つ文法が得られたことは、対象とした英語が右分岐の言語であると考えることの妥当性を傍証する。また、品詞列の独立性と分布の類似性から名詞句などの概念が得られたことは、自然言語を説明する文法におけるこれらの概念の必要性の根拠を与える。また、従来の統語情報を用いた文法推論に対し、品詞情報のみを用いて統語規則を獲得し、比較的高い精度が得られた。これは、作成コストがより低い品詞情報からの統語規則獲得が有効であることを意味する。

謝 辞

本研究を進めるにあたり、終始御指導下さいました長尾眞教授に心から御礼申し上げます。

さらに、本研究に関し助言を頂いた黒橋禎夫助手に感謝いたします。

最後に、本研究およびプログラミングに関して援助して下さった長尾研究室の皆様に感謝します。

参 考 文 献

1) C. E. Shannon, A Mathematical Theory of Communication, *Bell System Tech. J.*, Vol. 27, (1948), pp. 379–423, 623–656.

2) Mitch Marcus, Very Large Annotated Database of American English, *Proceedings of the DARPA Speech and Natural Language Workshop*, (1991), p. 430.

3) Mitchell P. Marcus and Beatrice Santorini, Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, Vol. 19, No. 2, (1993), pp. 313–330.

4) Mitchell Marcus, Grace Kim, and Mary Ann Marcinkiewicz, The Penn Treebank: A Revised Corpus Design for Extracting Predicate Argument Structure, *Proceedings of the International Symposium on Spoken Dialogue*, (1993), pp. 77–81.

5) Makoto Nagao and Shinsuke Mori, A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, *Proceedings of the 15th International Conference on Computational Linguistics*, (1994), pp. 611–615.

6) Noam Chomsky, Three Models for the Description of Language, *IRE Transaction on Infomation Theory*, Vol. 2, No. 3, (1956), pp. 113–124.

7) A. W. Biermann, An Interactive Finite-State Language Learner, *First USA-JAPAN Computer Conference*, (1972), pp. 13–20.

8) Bruce Knobe and Kathleen Knobe, A Method for Inferring Context-free Grammars, *Infomation and Controle*, (1976).

9) King-Sun Fu and Taylor L. Booth, Grammartical Inference: Introduction and Survey–Part I, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 8, No. 3, (1986), pp. 343–359.

10) King-Sun Fu and Taylor L. Booth, Grammartical Inference: Introduction and Survey–Part II, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 8, No. 3, (1986), pp. 360–375.

11) Fernando Pereira and Yves Schabes, Inside-Outside Reestimation from Partially Bracketed Corpora, *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, (1992), pp. 128–135.

12) Yves Schabes, Michal Roth, and Randy Osborne, Parsing the Wall Street Journal with the Inside-Outside Algorithm, *Proceedings of the Sixth European Chapter of the Association for Computational Linguistics*, (1993), pp. 341–347.

13) Eric Brill, Automatic Grammar Induction and Parsing Free Text: A Transformation-Based Approach, *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, (1993), pp. 259–265.

14) Kenneth Ward Church, A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text, *Proceedings of the Second Conference on Applied Natural Language Processing*, (1988), pp. 136–143.

15) Eric Brill, A Simple Rule-Based Part of Speech Tagger, *Proceedings of the Third Conference on Applied Natural Language Processing*, (1992), pp. 152–154.

16) Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun, A Practical Part-of-Speech Tagger, *Proceedings of the Third Conference on Applied Natural Language Processing*, (1992), pp. 133–140.

17) Eric Brill, Some Advances in Transformation-Based Part of Speech Tagging, *Proceedings of the Twelveth National Conference on Artificial Intelligence*, (1994), pp. 722–727.

18) E. Black, Parsing english by computer: The State of the Art, *Proceedings of the International Symposium on Spoken Dialogue*, (1993), pp. 77–81.

19) John E. Hopcroft and Jeffrey D. Ulman, *Introduction to Automata Theory, Language and Computation*, (Addison-Wesley Publishing, 1979).

20) E. M. Gold, Language Identification in the Limit, *Infomation and Controle*, (1967), pp. 447–474.

21) D. Angluin, Learning Regular Sets from Queries and Counterexamples, *Infomation and Computation*, Vol. 75, (1987), pp. 87–106.

22) Charles T. Hemphill, John J. Godfrey, and George R. Doddington, The ATIS Spoken Language Systems Pilot Corpus, *Proceedings of the DARPA Speech and Natural Language Workshop*, (1990), pp. 96–101.

23) Donald Hindle, *User Manual for Fidditch*, (1983).

24) D. Hindle and M. Rooth, Acquiring Disambiguation Rules from Text, *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, (1989).

25) Fernando Pereira, Naftali Tishiby, and Lillian Lee, Distributional Clustering of English Words, *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, (1993), pp. 183–190.

26) Philip Resnik, A Class-Based Approach to Lexical Discovery, *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, (1992), pp. 327–329.

27) Donald Hindle, Noun Classification from Predicate-Argument Structures, *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, (1990), pp. 268–275.

28) Michael R. Brent and Robert C. Berwick, Automatic Acquisition of Subcategorization Frames from Tagged Text, *Proceedings of the DARPA Speech and Natural Language Workshop*, (1991), pp. 342–345.

29) Michael R. Brent, Automatic Acquisition of Subcategorization Frames from Untagged Text, *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, (1991), pp. 152–155.

30) Chiristopher D. Manning, Automatic Acquisition of a Large Subcategorization Dictionary from Corpora, *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, (1993), pp. 235–242.

31) Eric Brill and Mitchell Marcus, Automatically Acquiring Phrase Structure Using Distributional Analysis, *Proceedings of the DARPA Speech and Natural Language Workshop*, (1992), pp. 155–159.

32) Eric Brill, David Magerman, Mitchell Marcus, and Beatrice Santorini, Deducing Linguistic Structure from the Statistics of Large Corpora, *Proceedings of the DARPA Speech and Natural Language Workshop*, (1990), pp. 275–282.

33) David M. Magerman and Mitchell P. Marcus, Parsing a Natural Language Using Mutual Information Statistics, *Proceedings of the Eighth National Conference on Artificial Intelligence*, (1990), pp. 984–989.

34) B. B. Greene and G. M. Rubin, Automatic Grammatical Tagging of English, Technical report, (Brown University, 1971).

35) E. Black, S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski, A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars, *Proceedings of the DARPA Speech and Natural Language Workshop*, (1991), pp. 306–311.

36) Masaki KIYONO and Jun ichi TSUJII, Linguistic Knowledge Acquisition from Parsing Failures, *Proceedings of the Sixth European Chapter of the Association for Computational Linguistics*, (1993), pp. 222–231.

37) Rey-Long Liu and Von Wun Soo, A Corpus-Based Learning Technique for Building a Self-Extensible Parser, *Proceedings of the 15th International Conference on Computational Linguistics*, (1994), pp. 441–446.

38) D. Hindle and M. Rooth, Structural Ambiguity and Lexical Relations, *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, (1991), pp. 229–236.

39) Claire Cardie, Corpus-Based Acquisition of Relative Pronoun Disambiguation Heuristics, *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, (1992), pp. 216–223.

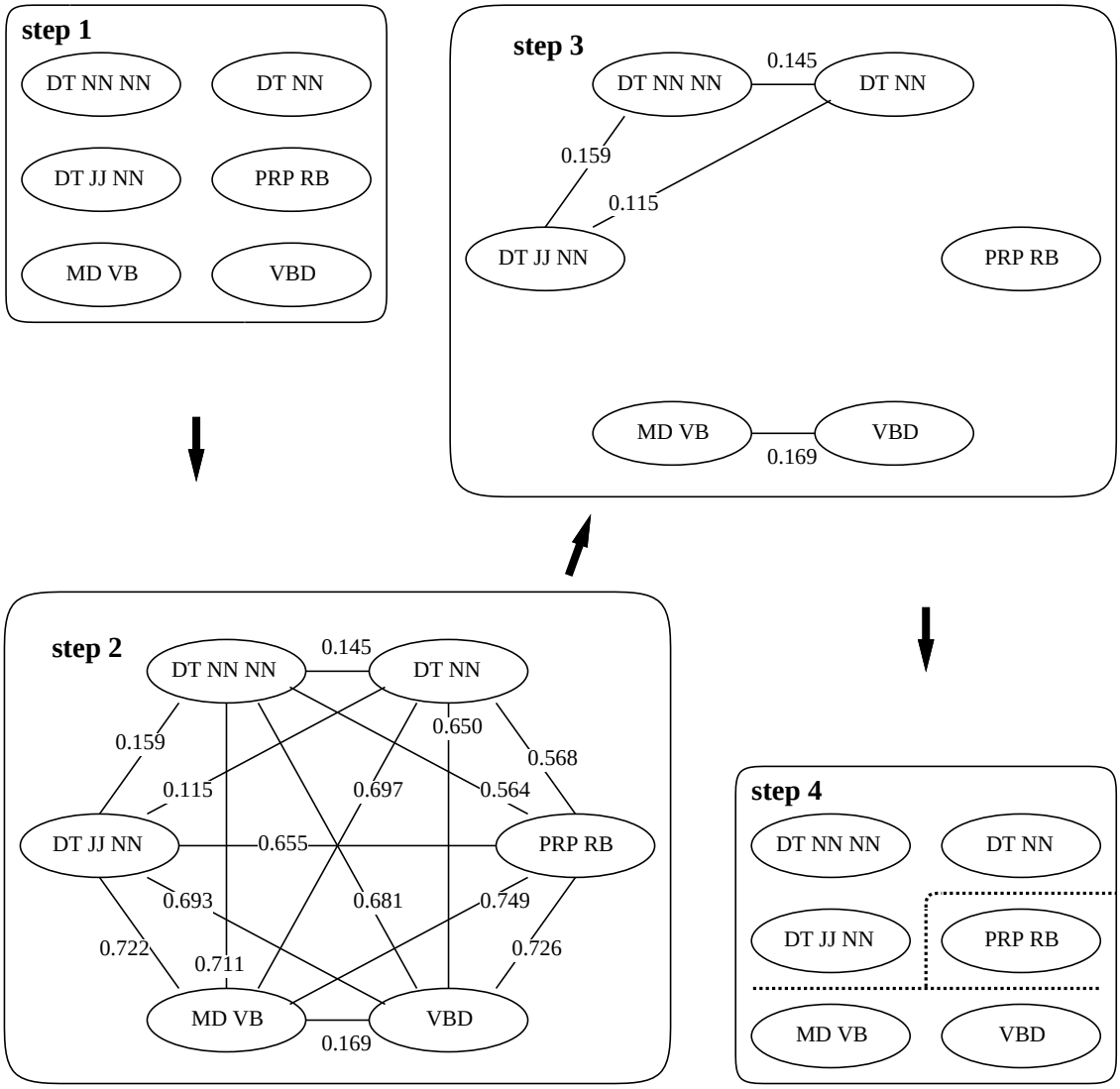

図 3.1: 品詞列のクラスタリング ($D_{min} = 0.25$)

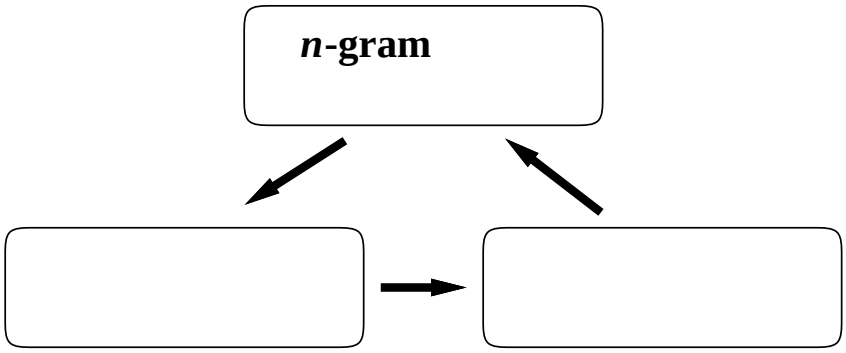

図 3.2: 処理の概念図

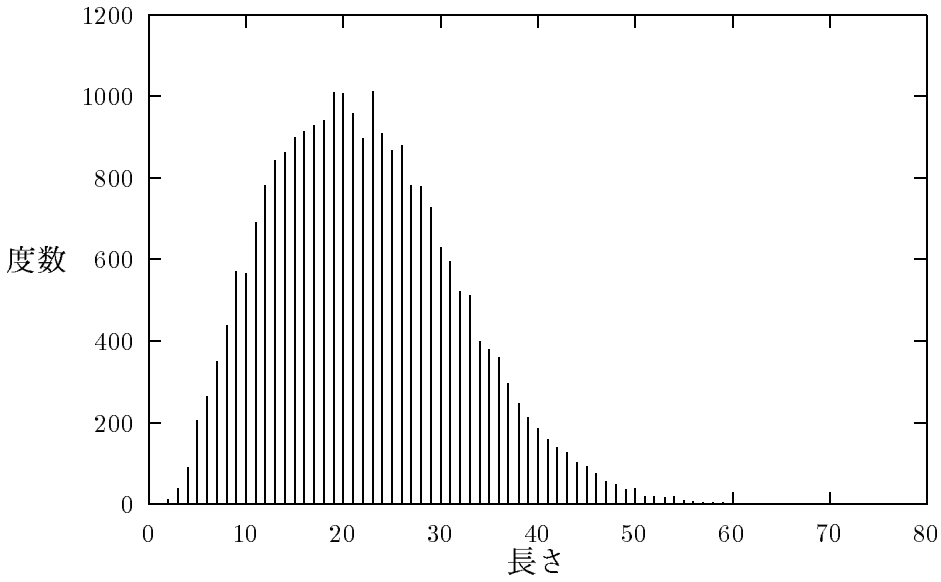

図 4.1: Wall Street Journal における文の長さの度数分布

```
文     :  This is a pen .
右線形  :  ( ( This ( is ( a pen ) ) ) .  )
左線形  :  ( ( ( ( This is ) a ) pen ) .  )
```

図 4.2: 右線形および左線形の文構造

表 2.1: Penn Treebank の品詞タグ一覧

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb
37.	,	Comma
38.	.	Sentence-final punctuation

表 2.2: Penn Treebank の統語タグ一覧

1.	ADJP	Adjective phrase
2.	ADVP	Adverb phrase
3.	NP	Noun phrase
4.	PP	Prepositional phrase
5.	S	Simple declarative clause
6.	SBAR	Clause introduced by a (possibly empty) subordinating conjunction
7.	SBARQ	Direct question introduced by a *wh*-word or *wh*-phrase
8.	SINV	Inverted declarative sentence
9.	SQ	That part of an SBARQ that excludes the *wh*-word or *wh*-phrase
10.	V P	Verb phrase
11.	WHADVP	*Wh*-adverb phrase
12.	WHNP	*Wh*-noun phrase
13.	WHPP	*Wh*-prepositional phrase
14.	X	Constituent of unknown or uncertain type

Null elements

1.	*	A zero pronoun
2.	0	A zero complementizer
3.	NIL	Legitimate in pied-piping contexts
4.	T	Trace. Marks the position where a fronted *wh*-constituent is interpreted

表 2.3: 品詞列 DT NN の左右の確率分布

```
品　　詞　　列 : DT    NN
頻　　　度 :  23758 ============
左エントロピー : 3.1974 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
右エントロピー : 3.5425 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
-----------------------------------------------------------------
                    =  2.86%   CC      2.70% =
                    =  0.14%   CD      0.25% =
                    =  0.12%   DT      0.62% =
                       0.00%   EX      0.00%
                       0.00%   FW      0.00%
               ========== 39.05%   IN     29.11% ========
                    =  0.33%   JJ      0.75% =
                    =  0.05%   JJR     0.08% =
                    =  0.05%   JJS     0.01% =
                       0.00%   LS      0.00%
                    =  0.03%   MD      2.67% =
                    =  1.57%   NN     12.69% ====
                    =  0.25%   NNP     0.48% =
                       0.00%   NNPS    0.01% =
                    =  1.49%   NNS     2.13% =
                    =  0.32%   PDT     0.00%
                       0.00%   POS     5.68% ==
                    =  0.17%   PRP     0.48% =
                       0.00%   PRP$    0.03% =
                    =  1.73%   RB      2.65% =
                    =  0.48%   RBR     0.47% =
                       0.00%   RBS     0.02% =
                    =  0.53%   RP      0.07% =
                       0.00%   SYM     0.00%
                    =  3.12%   TO      4.02% ==
                       0.00%   UH      0.00%
                   ==  7.86%   VB      0.14% =
                   ==  7.04%   VBD     7.35% ==
                    =  3.80%   VBG     0.91% =
                    =  1.96%   VBN     1.08% =
                    =  1.61%   VBP     0.45% =
                    =  3.97%   VBZ     6.57% ==
                    =  0.17%   WDT     1.08% =
                    =  0.09%   WP      0.16% =
                       0.00%   WP$     0.03% =
                    =  0.75%   WRB     0.29% =
                   ==  7.53%   ,       8.12% ===
                 ==== 12.89%   .       8.91% ===
```

表 2.4: 品詞列 DT JJ NN の左右の確率分布

```
品　　詞　　列 : DT    JJ    NN
頻　　　　度 :   7487 ====
左エントロピー : 3.0028 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
右エントロピー : 3.3004 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
-----------------------------------------------------------------
                    =  3.07%   CC      3.21% =
                    =  0.08%   CD      0.12% =
                    =  0.08%   DT      0.73% =
                       0.00%   EX      0.00%
                       0.00%   FW      0.00%
          ========== 43.85%   IN     32.63% =========
                    =  0.27%   JJ      0.63% =
                       0.00%   JJR     0.04% =
                    =  0.11%   JJS     0.00%
                       0.00%   LS      0.00%
                       0.00%   MD      2.32% =
                    =  0.60%   NN     13.21% ====
                    =  0.31%   NNP     0.47% =
                       0.00%   NNPS    0.00%
                    =  0.53%   NNS     2.08% =
                    =  0.16%   PDT     0.00%
                       0.00%   POS     1.12% =
                    =  0.16%   PRP     0.37% =
                       0.00%   PRP$    0.00%
                    =  1.88%   RB      1.44% =
                       0.00%   RBR     0.04% =
                       0.00%   RBS     0.00%
                    =  0.49%   RP      0.04% =
                       0.00%   SYM     0.00%
                    =  3.31%   TO      4.15% ==
                       0.00%   UH      0.00%
                  ===  8.25%   VB      0.15% =
                   ==  7.79%   VBD     3.83% =
                    =  3.95%   VBG     1.02% =
                    =  2.38%   VBN     1.46% =
                    =  1.82%   VBP     0.40% =
                   ==  5.08%   VBZ     3.91% =
                    =  0.04%   WDT     1.18% =
                       0.00%   WP      0.28% =
                       0.00%   WP$     0.00%
                    =  0.45%   WRB     0.36% =
                   ==  6.61%   ,      11.18% ===
                  ===  8.70%   .      13.57% ====
```

表 2.5: 品詞列 MD VB の左右の確率分布

```
品     詞     列 : MD    VB
頻           度 :   4926 ===
左エントロピー : 3.0286 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
右エントロピー : 3.7114 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
-----------------------------------------------------------------
                    =  2.27%   CC      0.47% =
                    =  0.55%   CD      1.95% =
                    =  1.87%   DT     19.31% =====
                    =  0.47%   EX      0.00%
                       0.00%   FW      0.00%
                    =  0.08%   IN     11.00% ===
                    =  0.28%   JJ      9.85% ===
                       0.00%   JJR     1.50% =
                       0.00%   JJS     0.00%
                       0.00%   LS      0.00%
                       0.00%   MD      0.06% =
              ======= 26.65%   NN      5.18% ==
                  === 11.43%   NNP     3.57% =
                    =  0.47%   NNPS    0.04% =
                ===== 19.37%   NNS     3.45% =
                       0.00%   PDT     0.06% =
                    =  0.08%   POS     0.00%
                ===== 16.87%   PRP     2.17% =
                       0.00%   PRP$    2.84% =
                    =  3.71%   RB      5.20% ==
                    =  0.04%   RBR     0.69% =
                       0.00%   RBS     0.04% =
                       0.00%   RP      1.18% =
                       0.00%   SYM     0.00%
                       0.00%   TO      6.96% ==
                       0.00%   UH      0.00%
                    =  0.16%   VB      0.69% =
                    =  0.12%   VBD     0.14% =
                    =  0.04%   VBG     2.52% =
                    =  0.26%   VBN    17.40% =====
                    =  0.16%   VBP     0.04% =
                    =  0.14%   VBZ     0.16% =
                   ==  6.68%   WDT     0.00%
                    =  1.44%   WP      0.06% =
                       0.00%   WP$     0.00%
                       0.00%   WRB     0.22% =
                   ==  6.76%   ,       1.02% =
                       0.00%   .       2.21% =
```

表 2.6: 品詞列の左右のエントロピー

$H_l + H_r$	f	H_l	H_r	V_l	V_r	$\boldsymbol{pos}$
8.42	27666	4.30	4.12	0.07	0.06	RB
7.87	1942	4.14	3.73	0.12	0.07	$RB \cdot RB$
7.61	11920	3.75	3.86	0.11	0.10	$TO \cdot VB$
7.56	1383	3.77	3.79	0.10	0.10	$NNS \cdot TO \cdot VB$
7.56	1359	3.79	3.76	0.10	0.12	$IN \cdot RB$
7.55	13598	3.80	3.74	0.10	0.10	VBG
7.54	1351	3.64	3.90	0.10	0.11	$PRP \cdot MD \cdot VB$
7.48	4548	3.53	3.94	0.08	0.12	$NNS \cdot VBD$
7.46	1077	3.21	4.25	0.07	0.16	$CC \cdot RB$
7.45	2128	3.64	3.81	0.11	0.14	$IN \cdot DT \cdot NNS$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.99	8450	1.35	2.64	0.25	0.46	POS
3.95	1067	1.73	2.22	0.28	0.46	$VB \cdot PRP\$$
3.75	1269	1.97	1.77	0.50	0.36	$VB \cdot DT \cdot JJ$
3.71	1902	0.56	3.14	0.13	0.84	$CD \cdot CD \cdot IN$
3.59	2183	2.86	0.73	0.74	0.23	$NN \cdot MD$
3.54	1394	3.05	0.48	0.87	0.16	$MD \cdot RB$
3.30	1059	0.86	2.44	0.27	0.70	$CD \cdot DT \cdot NN$
3.01	1779	1.24	1.77	0.43	0.48	$POS \cdot JJ$
2.95	1089	1.93	1.02	0.73	0.48	$NNS \cdot DT$
1.65	1150	1.01	0.63	0.85	0.66	$CD \cdot DT$

表 2.7: 品詞列の左右の分散

$V_l + V_r$	f	H_l	H_r	V_l	V_r	$\boldsymbol{pos}$
0.13	27666	4.30	4.12	0.07	0.06	RB
0.19	1942	4.14	3.73	0.12	0.07	$RB \cdot RB$
0.20	1383	3.77	3.79	0.10	0.10	$NNS \cdot TO \cdot VB$
0.20	2972	3.60	3.81	0.09	0.11	$PRP \cdot VBZ$
0.20	4548	3.53	3.94	0.08	0.12	$NNS \cdot VBD$
0.20	13598	3.80	3.74	0.10	0.10	VBG
0.21	1351	3.64	3.90	0.10	0.11	$PRP \cdot MD \cdot VB$
0.21	1359	3.79	3.76	0.10	0.12	$IN \cdot RB$
0.21	3736	3.44	3.92	0.09	0.12	$PRP \cdot VBD$
0.21	11920	3.75	3.86	0.11	0.10	$TO \cdot VB$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.85	8732	3.14	0.86	0.70	0.16	MD
0.86	1269	1.97	1.77	0.50	0.36	$VB \cdot DT \cdot JJ$
0.90	1077	2.52	1.54	0.61	0.28	$VBZ \cdot NNP$
0.92	1779	1.24	1.77	0.43	0.48	$POS \cdot JJ$
0.97	1059	0.86	2.44	0.27	0.70	$CD \cdot DT \cdot NN$
0.97	1902	0.56	3.14	0.13	0.84	$CD \cdot CD \cdot IN$
0.98	2183	2.86	0.73	0.74	0.23	$NN \cdot MD$
1.03	1394	3.05	0.48	0.87	0.16	$MD \cdot RB$
1.21	1089	1.93	1.02	0.73	0.48	$NNS \cdot DT$
1.51	1150	1.01	0.63	0.85	0.66	$CD \cdot DT$

表 2.8: 品詞列 DT NN との確率分布ベクトルの距離

d	f	$\boldsymbol{pos}$
0.115	7487	$DT \cdot JJ \cdot NN$
0.145	3015	$DT \cdot NN \cdot NN$
0.150	1102	$DT \cdot JJ \cdot NNS$
0.155	1106	$DT \cdot NN \cdot IN \cdot DT \cdot NN$
0.182	2137	$PRP\$ \cdot NN$
0.205	3905	$DT \cdot NNS$
0.215	1418	$DT \cdot NNP \cdot NN$
0.237	1026	$PRP\$ \cdot NNS$
0.308	4104	$CD \cdot NN$
0.314	9519	$JJ \cdot NNS$
0.339	1676	$JJ \cdot NN \cdot NNS$
0.340	1054	$JJ \cdot JJ \cdot NNS$
0.340	3333	$CD \cdot NNS$
0.363	1266	$VBG \cdot NN$
0.402	9155	VBG
0.417	2085	JJR
0.425	1724	$RB \cdot JJ$
0.433	7296	$NN \cdot NNS$
0.442	1374	$NNP \cdot POS \cdot NN$
0.444	2173	$PRP \cdot VBD$
⋮	⋮	⋮

表 3.1: 独立な品詞列の例

f	H_l	H_r	Pd_l	Pd_r	$\boldsymbol{pos}$
54724	3.05	3.26	0.164	0.187	NNP
39375	3.58	3.62	0.052	0.270	NNS
23758	3.20	3.54	0.204	0.170	$DT \cdot NN$
20088	3.25	3.36	0.236	0.254	$NNP \cdot NNP$
19055	3.10	3.83	0.091	0.061	VBD
18460	4.19	4.01	0.172	0.155	RB
14550	3.65	3.10	0.059	0.168	CD
13327	3.57	3.36	0.057	0.084	VBN
9519	3.77	3.47	0.081	0.336	$JJ \cdot NNS$
9385	3.15	3.21	0.072	0.219	$IN \cdot NNP$
9278	3.39	3.50	0.090	0.235	$IN \cdot DT \cdot NN$
7753	3.45	3.01	0.326	0.051	PRP
7487	3.00	3.30	0.153	0.247	$DT \cdot JJ \cdot NN$
7296	3.75	3.53	0.089	0.332	$NN \cdot NNS$
6553	3.21	3.40	0.114	0.259	$IN \cdot NN$
5313	3.16	3.14	0.214	0.325	$NNP \cdot NNP \cdot NNP$
4018	3.69	3.15	0.074	0.189	$IN \cdot CD$
3905	3.02	3.66	0.254	0.175	$DT \cdot NNS$
3722	3.10	3.63	0.068	0.218	$IN \cdot NNS$
3465	3.47	3.61	0.125	0.051	$NNS \cdot VBP$
⋮	⋮	⋮	⋮	⋮	⋮

表 3.2: クラスタリング結果の例

出現合計	f	H_l	H_r	$\boldsymbol{pos}$
20260	19055	3.09	3.83	VBD
	1205	3.16	3.82	$RB \cdot VBD$
86473	23758	3.19	3.54	$DT \cdot NN$
	3905	3.01	3.66	$DT \cdot NNS$
	2137	3.07	3.47	$PRP\$ \cdot NN$
	7487	3.00	3.30	$DT \cdot JJ \cdot NN$
	1102	3.08	3.66	$DT \cdot JJ \cdot NNS$
	1106	3.18	3.51	$DT \cdot NN \cdot IN \cdot DT \cdot NN$
80317	6553	3.21	3.40	$IN \cdot NN$
	3722	3.09	3.62	$IN \cdot NNS$
	9278	3.38	3.49	$IN \cdot DT \cdot NN$
	1499	3.56	3.68	$IN \cdot DT \cdot NNS$
	1771	3.06	3.30	$IN \cdot JJ \cdot NN$
	2555	3.22	3.31	$IN \cdot JJ \cdot NNS$
	3283	3.36	3.28	$IN \cdot DT \cdot JJ \cdot NN$
	1326	3.04	3.52	$IN \cdot DT \cdot NN \cdot NN$
3944	2173	3.04	3.77	$PRP \cdot VBD$
	1771	3.14	3.71	$PRP \cdot VBZ$
⋮	⋮	⋮	⋮	⋮

表 3.3: パラメーター

f_{min} : 抽出の対象とする品詞列の頻度の下限

H_{min} : 抽出の対象とする品詞列の左右のエントロピーの下限

Pd_{min} : 抽出の対象とする品詞列の左右のデリミターの割合の下限

D_{max} : 品詞列の類似度を表す確率分布ベクトルの距離の上限 (方法 2 のみ)

F : 統計結果の信頼性が限界となる出現頻度 (方法 2 のみ)

表 4.1: Penn Treebank の統語構造との比較

	生成規則数	非終端記号数	交差括弧数比
右線形	—	—	43.7%
左線形	—	—	75.7%
実験 1	490	490	35.7%
実験 2	956	19	27.0%
実験 3	594	25	12.7%

表 4.2: 実験 1 によって得られた生成規則の例

$NT001 \rightarrow DT \cdot NN$
$NT002 \rightarrow NNP \cdot NNP$
$NT003 \rightarrow DT \cdot JJ \cdot NN$
$NT004 \rightarrow JJ \cdot NNS$
$NT005 \rightarrow JJ \cdot NN$
$NT006 \rightarrow IN \cdot NT001$
$NT007 \rightarrow IN \cdot NN$
$NT008 \rightarrow IN \cdot NNP$
$NT009 \rightarrow NN \cdot NN$
$NT010 \rightarrow IN \cdot CD$
$NT011 \rightarrow DT \cdot NNS$
$NT012 \rightarrow IN \cdot NNS$
$NT013 \rightarrow NN \cdot NNS$
$NT014 \rightarrow IN \cdot NT002$
$NT015 \rightarrow IN \cdot NT003$
$NT016 \rightarrow IN \cdot DT \cdot NNP$
$NT017 \rightarrow IN \cdot NT004$
$NT018 \rightarrow CD \cdot NNS$
$NT019 \rightarrow NT002 \cdot NNP$
$NT020 \rightarrow PRP \cdot VBD$
$NT021 \rightarrow DT \cdot NNP$
$NT022 \rightarrow VBD \cdot VBN$
$NT023 \rightarrow PRP \cdot VBZ$
$NT024 \rightarrow IN \cdot NT005$
$NT025 \rightarrow IN \cdot DT \cdot JJ$
$NT026 \rightarrow PRP\$ \cdot NN$
$NT027 \rightarrow NT001 \cdot NN$
$NT028 \rightarrow NNS \cdot VBP$
$NT029 \rightarrow IN \cdot NT011$
$NT030 \rightarrow CC \cdot NNP$
$NT031 \rightarrow IN \cdot JJ$
$NT032 \rightarrow NNS \cdot VBD$
$NT033 \rightarrow DT \cdot JJ$
$NT034 \rightarrow RB \cdot RB$
$NT035 \rightarrow MD \cdot VB \cdot VBN$
$NT036 \rightarrow IN \cdot DT \cdot NT002$
$NT037 \rightarrow VBD \cdot NT001$
$NT038 \rightarrow NT006 \cdot NN$
$NT039 \rightarrow PRP \cdot MD \cdot VB$
$NT040 \rightarrow RB \cdot JJ$
$\vdots$
$\vdots$

表 4.3: 実験 2 によって得られた生成規則の例

$NT001 \rightarrow VBG \cdot NNS$
$NT001 \rightarrow PDT \cdot DT \cdot NN \cdot NNS$
$NT001 \rightarrow DT \cdot NN \cdot NN \cdot NNS$
$NT001 \rightarrow DT \cdot DT \cdot JJ \cdot NNS$
$NT002 \rightarrow TO \cdot NT001$
$NT002 \rightarrow TO \cdot VB \cdot NNS \cdot NNS$
$NT002 \rightarrow TO \cdot JJ \cdot NNS \cdot NNS$
$NT002 \rightarrow IN \cdot NT001$
$NT002 \rightarrow IN \cdot TO \cdot NT001$
$NT002 \rightarrow IN \cdot TO \cdot JJ \cdot NNS$
$NT003 \rightarrow NNP \cdot NNP \cdot NNP$
$NT003 \rightarrow NNP$
$NT004 \rightarrow NN$
$NT004 \rightarrow JJ \cdot NN$
$NT002 \rightarrow NT002 \cdot NT003$
$NT002 \rightarrow RB \cdot NT002$
$\vdots$

$NT001 \rightarrow PRP\$ \cdot JJ \cdot NN$
$NT001 \rightarrow DT \cdot PRP\$ \cdot NN$
$NT001 \rightarrow DT \cdot JJ \cdot NN$
$NT001 \rightarrow DT \cdot DT \cdot JJ \cdot NN$
$NT002 \rightarrow TO \cdot VB \cdot NT001$
$NT002 \rightarrow TO \cdot NNS \cdot NNS$
$NT002 \rightarrow RB \cdot TO \cdot NT001$
$NT002 \rightarrow IN \cdot VBG \cdot NT001$
$NT002 \rightarrow IN \cdot TO \cdot VB \cdot NT001$
$NT002 \rightarrow RB \cdot RB \cdot NT002 \cdot NT002$
$NT003 \rightarrow NNP \cdot NNP$
$NT004 \rightarrow NN \cdot NN$
$NT004 \rightarrow JJ \cdot NN \cdot NN$
$NT002 \rightarrow NT002 \cdot NT004$
$NT002 \rightarrow NT002 \cdot VBG \cdot NT002$
$NT002 \rightarrow IN \cdot NT003$
$\vdots$

表 4.4: 実験 3 によって得られた生成規則の例

$NT001$	$\rightarrow$	$VBG \cdot NN \cdot NNS$	$NT001$	$\rightarrow$	$PRP\$ \cdot NN$
$NT001$	$\rightarrow$	$PDT \cdot PRP\$ \cdot NN \cdot NNS$	$NT001$	$\rightarrow$	$PDT \cdot DT \cdot JJ \cdot NN$
$NT001$	$\rightarrow$	$DT \cdot VBN \cdot NN \cdot NNS$	$NT001$	$\rightarrow$	$DT \cdot NN \cdot NNS$
$NT002$	$\rightarrow$	$TO \cdot NT001$	$NT002$	$\rightarrow$	$IN \cdot NT001$
$NT002$	$\rightarrow$	$IN \cdot VBG \cdot NT001 \cdot NT001$	$NT002$	$\rightarrow$	$IN \cdot VBG \cdot NT001$
$NT002$	$\rightarrow$	$IN \cdot JJR \cdot NNS \cdot NNS$	$NT002$	$\rightarrow$	$IN \cdot IN \cdot NT001 \cdot RB$
$NT002$	$\rightarrow$	$IN \cdot CD \cdot JJ \cdot NNS \cdot NNS$	$NT002$	$\rightarrow$	$IN \cdot CD \cdot JJ \cdot NNS$
$NT003$	$\rightarrow$	$NN \cdot NT002$	$NT003$	$\rightarrow$	$NN \cdot NN \cdot NT002$
$NT003$	$\rightarrow$	$NN \cdot NN$	$NT003$	$\rightarrow$	NN
$NT003$	$\rightarrow$	$JJ \cdot NN \cdot NT002$	$NT003$	$\rightarrow$	$JJ \cdot NN \cdot NN \cdot NT002$
$NT003$	$\rightarrow$	$JJ \cdot JJ \cdot NN \cdot NN \cdot NT002$	$NT004$	$\rightarrow$	$NNP \cdot NNP \cdot NNP$
$NT004$	$\rightarrow$	$NNP \cdot NNP$	$NT004$	$\rightarrow$	$NNP \cdot$
$NT002$	$\rightarrow$	$NT002 \cdot NT002$	$NT002$	$\rightarrow$	$IN \cdot NT003$
$NT002$	$\rightarrow$	$IN \cdot IN \cdot NT002 \cdot NT002$	$NT002$	$\rightarrow$	$IN \cdot IN \cdot IN \cdot NT003$
	$\vdots$			$\vdots$	

表 4.5: 実験 3 による文法の一部

$NT001 \rightarrow PRP\$ \cdot NNS \cdot NNS$
$NT001 \rightarrow NNP \cdot NNP \cdot POS \cdot NN$
$NT001 \rightarrow DT \cdot JJ \cdot NN$
$NT002 \rightarrow TO \cdot VB \cdot NT001$
$NT002 \rightarrow RB \cdot TO \cdot NT001$
$NT002 \rightarrow IN \cdot RB \cdot NT001$
$NT002 \rightarrow TO \cdot NT003$
$NT002 \rightarrow RB \cdot IN \cdot NT003$
$NT003 \rightarrow NNP \cdot NNP \cdot NNP$
$NT003 \rightarrow NT003 \cdot CC \cdot NT003$
$NT004 \rightarrow JJ \cdot NN \cdot NN$
$NT005 \rightarrow NT001 \cdot NT002$
$NT005 \rightarrow NT001 \cdot VBG \cdot NT002$
$NT005 \rightarrow VBG \cdot NT001$
$NT006 \rightarrow VBZ \cdot NT005 \cdot NT005$
$NT006 \rightarrow VBZ \cdot NT005 \cdot NT001$
$NT006 \rightarrow VBZ \cdot NT002 \cdot NT005$
$NT006 \rightarrow VBZ \cdot NT001 \cdot NT001$
$NT006 \rightarrow VBD \cdot NT005 \cdot NT005$
$NT006 \rightarrow MD \cdot VB \cdot NT001$
$NT007 \rightarrow VBZ \cdot VBN \cdot VBN$
$NT008 \rightarrow NNS \cdot NT006$
$NT008 \rightarrow NNS \cdot VBP \cdot NT001$
$NT009 \rightarrow NT003 \cdot JJ \cdot NT008$
$NT010 \rightarrow VBG \cdot NT004 \cdot NT006$
$NT001 \rightarrow PDT \cdot DT \cdot NNS \cdot NN$
$NT001 \rightarrow DT \cdot VBG \cdot NNS \cdot NNS$
$NT002 \rightarrow TO \cdot NT001$
$NT002 \rightarrow TO \cdot VB \cdot NNS$
$NT002 \rightarrow IN \cdot NT001$
$NT002 \rightarrow IN \cdot NNS \cdot NN$
$NT002 \rightarrow RB \cdot NT002$
$NT002 \rightarrow IN \cdot NT003$
$NT003 \rightarrow NNP \cdot NNP$
$NT004 \rightarrow NN \cdot NN$
$NT005 \rightarrow NT002 \cdot VBN \cdot NT002$
$NT005 \rightarrow NT001 \cdot VBN \cdot NT002$
$NT005 \rightarrow VBN \cdot NT002$
$NT005 \rightarrow PRP \cdot MD \cdot VB \cdot NT002$
$NT006 \rightarrow VBZ \cdot NT005 \cdot NT002$
$NT006 \rightarrow VBZ \cdot NT005$
$NT006 \rightarrow VBZ \cdot NT002$
$NT006 \rightarrow VBZ \cdot NT001$
$NT006 \rightarrow VBD \cdot NT002$
$NT006 \rightarrow MD \cdot VB \cdot NT001$
$NT007 \rightarrow MD \cdot VB$
$NT008 \rightarrow NNS \cdot VBP \cdot NT005$
$NT009 \rightarrow NT004 \cdot NT008$
$NT010 \rightarrow WRB \cdot PRP \cdot NT006$
$NT010 \rightarrow VBG \cdot NT004$

表 4.6: 実験 3 によって導入された非終端記号の解釈

$NT001$	:	名詞句 (冠詞あり)	$NT006$	:	動詞句
$NT002$	:	前置詞句	$NT007$	:	動詞句 (目的語なし)
$NT003$	:	固有名詞句	$NT008$	:	文 (主語は NNS)
$NT004$	:	名詞句 (冠詞なし)	$NT009$	:	名詞句 (文による修飾がある)
$NT005$	:	目的語	$NT010$	:	目的語 (間接疑問文を含む)

付録 A 数式の変形

A.1 分散

式 2.10 と式 2.11 が同値であることを示す。

式 2.10 から

$$
\begin{aligned}
V &= \sum_{i=1}^{n}(P_i - \overline{P})^2 && \text{(A.1)}\\
&= \sum_{i=1}^{n}(P_i^2 - 2P_i\overline{P} + \overline{P}^2) && \text{(A.2)}\\
&= \sum_{i=1}^{n} P_i^2 - \sum_{i=1}^{n} 2P_i\overline{P} + \sum_{i=1}^{n}\overline{P}^2 && \text{(A.3)}\\
&= \sum_{i=1}^{n} P_i^2 - 2\overline{P}\sum_{i=1}^{n} P_i + \overline{P}^2\sum_{i=1}^{n} 1 && \text{(A.4)}\\
&= \sum_{i=1}^{n} P_i^2 - 2n\overline{P}^2 + n\overline{P}^2 \qquad (\overline{P} = \frac{1}{n}\sum_{i=1}^{n} P_i) && \text{(A.5)}\\
&= |\boldsymbol{P}|^2 - n\overline{P}^2 && \text{(A.6)}
\end{aligned}
$$

である。よって

$$V = \sum_{i=1}^{n}(P_i - \overline{P})^2 = |\boldsymbol{P}|^2 - n\overline{P}^2$$

である。

A.2 エントロピーの最大値

頻度を f として、等確率分布のときエントロピーは最大値 $H_{max} = \log f$ となることを示す。

$$\boldsymbol{P} = (P_1,\ P_2,\ \ldots,\ P_f) \quad (P_i = \frac{1}{f},\ i = 1,2,\ldots,f)$$

と表せることを用いて、

$$H_{max} = -\sum_{i=1}^{f} P_i\ log\ P_i \qquad \text{(A.7)}$$

$$= -\sum_{i=1}^{f} \frac{1}{f}\, log\, \frac{1}{f} \tag{A.8}$$

$$= -\frac{1}{f}\, log\, \frac{1}{f} \sum_{i=1}^{f} 1 \tag{A.9}$$

$$= \frac{1}{f} log f \cdot f \tag{A.10}$$

$$= \log f \tag{A.11}$$

である。よって

$$H_{max} = \log f$$

である。

A.3 分散の最小値

頻度を f として、等確率分布のとき分散は最小値 $V_{max} = 0$ となることを示す。

$$\boldsymbol{P} = (P_1,\, P_2,\, \ldots,\, P_f) \quad (P_i = \frac{1}{f},\, i = 1, 2, \ldots, f)$$

と表せることを用いて、

$$V_{min} = \sum_{i=1}^{f} (P_i - \overline{P})^2 \tag{A.12}$$

$$= \sum_{i=1}^{f} (\frac{1}{f} - \frac{1}{f}) \tag{A.13}$$

$$= 0 \tag{A.14}$$

である。よって

$$V_{min} = 0$$

である。

付録 B 解析結果の例

以下は、それぞれの方法によるコーパスの最終状態とあらかじめコーパスに付加された構造を最初の 20 文について比較した結果である。各文に付加されている情報は以下の 3 つである。

- 文のデータ
- 各実験において出力された構造 (NT で始まる記号は非終端記号である)
- Penn Treebank における構造

文のデータに含まれる情報は以下の 4 つから構成される。

number: 文番号
pare1 : 各実験において付加された括弧対の数
pare2 : Penn Treebank の構造における括弧対の数
cross : 交差括弧数

B.1 方法 1

```
number = 1, pare1 = 14, pare2 = 24, cross = 7

    ( NNP ( NT490 ( NT231 VBZ VBN ) ( NT001 DT NN ) ) ( NT098 (
    NT033 DT JJ ) ( NT005 JJ NN ) ) , ( NT001 DT NN ) ( NT002 NNP
    NNP ) ( NT116 VBZ ( NT062 RB VBN ) ) IN PRP$ CD ( NT111 JJR NNS
    ) , DT IN WDT ( NT022 VBD VBN ) ( NT029 IN ( NT011 DT NNS ) ) .
    )

    ( ( ( NNP ) VBZ ( VBN ( DT NN ) ( ( DT ( JJ JJ ) NN ) , ( ( DT
    NN ) ( ( NNP NNP ) VBZ RB ( VBN ( IN ( ( PRP$ CD JJR NNS ) , ( (
    DT ( IN ( WDT ) ) ) ( VBD ( VBN ( IN ( DT NNS ) ) ) ) ) ) ) ) )
    ) ) ) ) . )
```

number = 2, pare1 = 8, pare2 = 14, cross = 4

((NT061 IN RB) , (NT002 NNP NNP) (NT263 (NT070 VBZ VBG) TO VB) NNP (NT462 IN (NT150 VBG (NT003 DT JJ NN))) IN (NT336 VBG JJ) .)

((IN RB , ((NNP NNP) VBZ (VBG (TO (VB (NNP) (IN (VBG (DT JJ NN))) (IN (VBG (JJ))))))))) .)

number = 3, pare1 = 9, pare2 = 14, cross = 6

((NT364 CC (NT027 (NT001 DT NN) NN)) VBZ RP (NT389 (NT001 DT NN) (NT038 (NT006 IN (NT001 DT NN)) NN)) , (NT144 IN DT) DT VBZ .)

((CC ((DT NN NN) (VBZ RP (DT NN (IN (DT NN NN , (IN ((DT) ((DT) (VBZ))))))))))) .)

number = 4, pare1 = 19, pare2 = 27, cross = 9

((NT006 IN (NT001 DT NN)) (NT060 VBZ (NT001 DT NN)) , (NT007 IN NN) , (NT002 NNP NNP) (NT075 MD VB (NT001 DT NN)) (NT158 IN (NT056 VBG (NT001 DT NN))) (NT008 IN NNP) (NT288 CC VBG) (NT027 (NT001 DT NN) NN) (NT029 IN (NT011 DT NNS)) (NT006 IN (NT001 DT NN)) .)

((IN (((DT NN) (VBZ (DT NN))) , (IN (NN))) , ((NNP NNP) MD (VB (DT NN (IN ((VBG (DT NN) (IN (NNP))) CC (VBG ((DT NN) (NN (IN (DT NNS (IN (DT NN)))))))))))))) .)

number = 5, pare1 = 5, pare2 = 8, cross = 0

((NT139 PRP RB) (NT383 VBZ (NT027 (NT001 DT NN) NN)) ,
(NT028 NNS VBP) .)

(((PRP) RB (VBZ (DT NN NN)) , ((NNS) (VBP))) .)

number = 6, pare1 = 9, pare2 = 14, cross = 7

((NT028 NNS VBP) (NT002 NNP NNP) (NT229 VBD RP) (NT003
DT JJ NN) TO (NT110 NN (NT007 IN NN)) (NT447 (NT014 IN (
NT002 NNP NNP)) NNP) .)

(((NNS) (VBP ((NNP NNP) (VBD RP (DT JJ NN (TO (NN)
)) (IN (NN (IN (NNP NNP NNP))))))))) .)

number = 7, pare1 = 19, pare2 = 31, cross = 6

((NT056 VBG (NT001 DT NN)) (NT006 IN (NT001 DT NN)) (
NT010 IN CD) , (NT020 PRP VBD) (NT007 IN NN) , IN (NT060
VBZ (NT001 DT NN)) (NT012 IN NNS) , VBZ (NT019 (NT002 NNP
NNP) NNP) , (NT027 (NT001 DT NN) NN) WP (NT458 VBZ (
NT056 VBG (NT001 DT NN))) (NT014 IN (NT002 NNP NNP)) .)

(((VBG (DT NN) (IN (DT NN (IN (CD)))))) , ((PRP)
(VBD (IN (NN))) , (IN (VBZ (DT NN (IN (NNS)))))
) , ((VBZ) ((NNP NNP NNP) , ((DT NN NN) ((WP) (VBZ
(VBG (DT NN (IN (NNP NNP)))))))))) .)

number = 8, pare1 = 4, pare2 = 9, cross = 1

(CC (NT139 PRP RB) (NT037 VBD (NT001 DT NN)) TO VB (
NT003 DT JJ NN) .)

((CC ((PRP) RB (VBD (DT NN (TO (VB (DT JJ NN))))))
)) .)

number = 9, pare1 = 8, pare2 = 17, cross = 2

((NT010 IN CD) , (NT002 NNP NNP) VBD (NT166 (NT007 IN NN
) NNS) TO VB (NT013 NN NNS) WP (NT155 VBD VBG) (NT027 (
NT001 DT NN) NN) .)

(((IN (CD)) , ((NNP NNP) (VBD IN (NN NNS) (TO (VB
((NN NNS) ((WP) (VBD (VBG (DT NN NN))))))))))
.)

number = 10, pare1 = 11, pare2 = 21, cross = 1

(DT VBN (NT019 (NT002 NNP NNP) NNP) , (NT096 NNP POS NN)
(NT258 JJ (NT006 IN (NT001 DT NN))) , WP (NT002 NNP NNP)
(NT022 VBD VBN) IN (NT259 VBG (NT006 IN (NT001 DT NN)))
.)

(((DT) (VBN (((NNP NNP NNP) , ((NNP) POS NN (JJ) (
IN (DT NN)))) , ((WP) ((NNP NNP) VBD (VBN (IN (VBG
(IN (DT NN))))))))))) .)

number = 11, pare1 = 12, pare2 = 21, cross = 7

((NT046 CC (NT002 NNP NNP)) (NT064 VBD RB) (NT029 IN (
NT011 DT NNS)) , (NT150 VBG (NT003 DT JJ NN)) (NT104 IN (

NT023 PRP VBZ)) (NT020 PRP VBD) (NT228 (NT020 PRP VBD) RB) IN .)

((CC ((NNP NNP) (VBD RB (IN (DT NNS)) , (VBG ((DT JJ NN) (IN ((PRP) (VBZ ((PRP) (VBD ((PRP) (VBD)) (RB IN))))))))))) .)

number = 12, pare1 = 12, pare2 = 16, cross = 2

((NT094 (NT001 DT NN) (NT006 IN (NT001 DT NN))) , (NT002 NNP NNP) (NT205 VBD (NT005 JJ NN)) (NT270 IN (NT109 PRP$ (NT009 NN NN))) (NT318 VBD (NT034 RB RB)) IN PRP$ NNS .)

(((DT NN (IN (DT NN))) , ((NNP NNP) (VBD (JJ NN) (IN ((PRP$ NN NN) (VBD (RB RB (IN (PRP$ NNS)))))))))) .)

number = 13, pare1 = 12, pare2 = 20, cross = 4

(VBG (NT165 RB (NT007 IN NN)) , (NT020 PRP VBD) (NT019 (NT002 NNP NNP) NNP) (NT016 IN DT NNP) (NT008 IN NNP) (NT007 IN NN) CD , IN (NT216 VBG RB) (NT172 CD (NT018 CD NNS)) (NT007 IN NN) .)

(((VBG RB (IN (NN))) , ((PRP) (VBD (NNP NNP NNP) (IN (DT NNP (IN (NNP)))) (IN (NN CD)) , (IN (VBG (RB CD CD NNS (IN (NN))))))))) .)

number = 14, pare1 = 7, pare2 = 11, cross = 2

((NT020 PRP VBD) (NT026 PRP$ NN) (NT024 IN (NT005 JJ NN)) (NT036 IN DT (NT002 NNP NNP)) (NT008 IN NNP) .)

(((PRP) (VBD (PRP$ NN (IN (JJ NN)) (IN (DT NNP NNP (IN (NNP)))))))) .)

number = 15, pare1 = 6, pare2 = 7, cross = 2

((NT002 NNP NNP) (NT457 RB (NT231 VBZ VBN)) (NT285 (NT003 DT JJ NN) (NT008 IN NNP)) .)

(((NNP NNP) RB VBZ (VBN (DT JJ NN (IN (NNP))))) .)

number = 16, pare1 = 12, pare2 = 24, cross = 9

(VBG (NT061 IN RB) NN (NT005 JJ NN) (NT010 IN CD) , (NT268 (NT020 PRP VBD) (NT001 DT NN)) WDT (NT022 VBD VBN) (NT007 IN NN) CC (NT209 VBN (NT015 IN (NT003 DT JJ NN))) IN PRP$ (NT235 NNS NN) .)

(((VBG IN (RB (NN JJ NN)) (IN (CD))) , ((PRP) (VBD ((DT NN) ((WDT) (VBD ((VBN (IN (NN))) CC (VBN (IN (DT JJ NN (IN (PRP$ NNS NN))))))))))))) .)

number = 17, pare1 = 7, pare2 = 10, cross = 4

((NT002 NNP NNP) VBD (NT127 VBG (NT004 JJ NNS)) , (NT296 JJ (NT007 IN NN)) (NT058 CC NN) (NT052 CC NNS) .)

(((NNP NNP) (VBD (VBG ((JJ NNS) , (JJ (IN (NN CC NN CC NNS))))))) .)

number = 18, pare1 = 4, pare2 = 10, cross = 1

(PRP (NT047 RB VBD) (NT044 CD NN) WDT VBD (NT224 (NT009 NN NN) NNS) .)

(((PRP) RB (VBD ((CD NN) ((WDT) (VBD (NN NN NNS)))))) .)

number = 19, pare1 = 19, pare2 = 22, cross = 10

((NT015 IN (NT003 DT JJ NN)) , (NT020 PRP VBD) (NT240 VBG RP) (NT001 DT NN) (NT090 CC (NT009 NN NN)) (NT480 (NT006 IN (NT001 DT NN)) (NT009 NN NN)) (NT013 NN NNS) (NT007 IN NN) (NT219 NN (NT058 CC NN)) , (NT400 RB (NT012 IN NNS)) (NT024 IN (NT005 JJ NN)) .)

(((IN (DT JJ NN)) , ((PRP) (VBD (VBG RP (DT NN CC NN NN (IN (DT NN))) (NN (NN (NN ((NNS (IN (NN NN CC NN))) , (RB IN (NNS (IN (JJ NN)))))))))))))) .)

number = 20, pare1 = 6, pare2 = 11, cross = 3

((NT001 DT NN) (NT057 VBD TO VB) (NT013 NN NNS) (NT012 IN NNS) (NT024 IN (NT005 JJ NN)) .)

(((DT NN) (VBD (TO (VB (NN NNS) (IN (NNS (IN (JJ NN)))))))) .)

B.2 方法 2

number = 1, pare1 = 14, pare2 = 24, cross = 3

((NT003 NNP) (NT006 VBZ VBN (NT001 DT NN) (NT001 DT JJ JJ NN)) , (NT001 DT NN) (NT003 NNP NNP) (NT016 (NT012 (NT007 VBZ) RB) VBN) IN PRP$ CD (NT001 JJR NNS) , DT IN WDT (NT006 VBD (NT005 VBN (NT002 IN (NT001 DT NNS)))) .)

(((NNP) VBZ (VBN (DT NN) ((DT (JJ JJ) NN) , ((DT NN) ((NNP NNP) VBZ RB (VBN (IN ((PRP$ CD JJR NNS) , ((DT (IN (WDT))) (VBD (VBN (IN (DT NNS)))))))))))))) .)

number = 2, pare1 = 9, pare2 = 14, cross = 12

(IN RB , (NT003 NNP NNP) (NT016 (NT007 (NT006 (NT007 (NT007 VBZ VBG) TO VB) (NT003 NNP) (NT002 IN VBG (NT001 DT JJ NN))) IN VBG) JJ) .)

((IN RB , ((NNP NNP) VBZ (VBG (TO (VB (NNP) (IN (VBG (DT JJ NN))) (IN (VBG (JJ)))))))) .)

number = 3, pare1 = 8, pare2 = 14, cross = 6

(CC (NT001 DT NN NN) (NT006 (NT007 VBZ) RP (NT005 (NT001 DT NN) (NT002 IN (NT001 DT NN NN)))) , IN DT DT (NT007 VBZ) .)

((CC ((DT NN NN) (VBZ RP (DT NN (IN (DT NN NN , (IN ((DT) ((DT) (VBZ)))))))))) .)

number = 4, pare1 = 20, pare2 = 27, cross = 13

((NT002 IN (NT001 DT NN)) (NT006 VBZ (NT001 DT NN)) , (NT002 IN NN) , (NT003 NNP NNP) (NT006 MD VB (NT005 (NT001 DT NN) (NT002 (NT002 IN VBG (NT001 DT NN)) IN (NT003 NNP)))) CC VBG (NT005 (NT001 DT NN NN) (NT002 (NT002 IN (NT001 DT NNS)) (NT002 IN (NT001 DT NN)))) .)

((IN (((DT NN) (VBZ (DT NN))) , (IN (NN))) , ((NNP NNP) MD (VB (DT NN (IN ((VBG (DT NN) (IN (NNP))) CC (VBG ((DT NN) (NN (IN (DT NNS (IN (DT NN)))))))))))))) .)

number = 5, pare1 = 4, pare2 = 8, cross = 0

(PRP (NT006 RB (NT006 VBZ (NT001 DT NN NN))) , (NT008 NNS) VBP .)

(((PRP) RB (VBZ (DT NN NN)) , ((NNS) (VBP))) .)

number = 6, pare1 = 9, pare2 = 14, cross = 2

((NT008 NNS) VBP (NT003 NNP NNP) (NT006 (NT007 VBD) RP (NT001 DT JJ NN)) TO (NT004 NN) (NT002 (NT002 IN NN) IN (NT003 NNP NNP NNP)) .)

(((NNS) (VBP ((NNP NNP) (VBD RP (DT JJ NN (TO (NN))) (IN (NN (IN (NNP NNP NNP))))))))) .)

number = 7, pare1 = 19, pare2 = 31, cross = 3

(VBG (NT005 (NT001 DT NN) (NT002 IN (NT001 DT NN))) IN CD , PRP (NT006 VBD (NT002 IN NN)) , (NT011 IN (NT006 VBZ (NT005 (NT001 DT NN) (NT002 IN NNS)))) , (NT007 VBZ) (NT003 NNP NNP NNP) , (NT001 DT NN NN) WP (NT006 VBZ VBG (NT005 (NT001 DT NN) (NT002 IN (NT003 NNP NNP)))) .)

(((VBG (DT NN) (IN (DT NN (IN (CD))))) , ((PRP) (VBD (IN (NN))) , (IN (VBZ (DT NN (IN (NNS))))))) , ((VBZ) ((NNP NNP NNP) , ((DT NN NN) ((WP) (VBZ (VBG (DT NN (IN (NNP NNP))))))))))) .)

number = 8, pare1 = 6, pare2 = 9, cross = 0

(CC PRP (NT006 RB (NT006 VBD (NT005 (NT001 DT NN) (NT002 TO VB (NT001 DT JJ NN))))) .)

((CC ((PRP) RB (VBD (DT NN (TO (VB (DT JJ NN))))))) .)

number = 9, pare1 = 8, pare2 = 17, cross = 8

(IN CD , (NT003 NNP NNP) (NT006 VBD (NT002 (NT002 IN NN NNS) (NT002 TO VB NN NNS))) WP (NT006 VBD (NT005 VBG (NT001 DT NN NN))) .)

(((IN (CD)) , ((NNP NNP) (VBD IN (NN NNS) (TO (VB ((NN NNS) ((WP) (VBD (VBG (DT NN NN))))))))))) .)

number = 10, pare1 = 12, pare2 = 21, cross = 0

(DT VBN (NT003 NNP NNP NNP) , (NT003 NNP) POS (NT004 NN) (NT005 JJ (NT002 IN (NT001 DT NN))) , WP (NT003 NNP NNP) (NT006 VBD (NT005 VBN (NT002 IN VBG (NT002 IN (NT001 DT NN))))) .)

(((DT) (VBN (((NNP NNP NNP) , ((NNP) POS NN (JJ) (IN (DT NN)))) , ((WP) ((NNP NNP) VBD (VBN (IN (VBG (IN (DT NN)))))))))) .)

number = 11, pare1 = 11, pare2 = 21, cross = 3

(CC (NT003 NNP NNP) (NT006 VBD (NT002 RB (NT002 IN (NT001 DT NNS)))) , (NT005 VBG (NT001 DT JJ NN)) IN PRP (NT007 VBZ) PRP (NT007 VBD) PRP (NT012 (NT007 VBD) RB) IN .)

((CC ((NNP NNP) (VBD RB (IN (DT NNS)) , (VBG ((DT JJ NN) (IN ((PRP) (VBZ ((PRP) (VBD ((PRP) (VBD)) (RB IN))))))))))) .)

number = 12, pare1 = 15, pare2 = 16, cross = 2

((NT005 (NT001 DT NN) (NT002 IN (NT001 DT NN))) , (NT003 NNP NNP) (NT016 (NT007 VBD) JJ) (NT004 NN) (NT002 IN (NT001 PRP$ NN NN)) (NT006 VBD (NT002 RB (NT002 RB (NT002 IN (NT001 PRP$ NNS))))) .)

(((DT NN (IN (DT NN))) , ((NNP NNP) (VBD (JJ NN) (IN ((PRP$ NN NN) (VBD (RB RB (IN (PRP$ NNS))))))))) .)

number = 13, pare1 = 16, pare2 = 20, cross = 6

((NT005 VBG (NT002 RB (NT002 IN NN))) , PRP (NT006 VBD (NT003 NNP NNP NNP) (NT002 (NT002 (NT002 IN DT (NT003 NNP)) (NT002 IN (NT003 NNP))) (NT002 IN NN))) CD , IN VBG RB (NT010 CD (NT010 CD (NT008 NNS (NT002 IN NN)))) .)

(((VBG RB (IN (NN))) , ((PRP) (VBD (NNP NNP NNP) (IN (DT NNP (IN (NNP)))) (IN (NN CD)) , (IN (VBG (RB CD CD NNS (IN (NN)))))))) .)

number = 14, pare1 = 9, pare2 = 11, cross = 2

(PRP (NT006 VBD (NT005 (NT001 PRP$ NN) (NT002 (NT002 (NT002 IN JJ NN) IN DT (NT003 NNP NNP)) (NT002 IN (NT003 NNP))))) .)

(((PRP) (VBD (PRP$ NN (IN (JJ NN)) (IN (DT NNP NNP (IN (NNP))))))) .)

number = 15, pare1 = 7, pare2 = 7, cross = 0

((NT003 NNP NNP) (NT006 RB (NT006 VBZ VBN (NT005 (NT001 DT JJ NN) (NT002 IN (NT003 NNP))))) .)

(((NNP NNP) RB VBZ (VBN (DT JJ NN (IN (NNP))))) .)

number = 16, pare1 = 16, pare2 = 24, cross = 3

(VBG IN RB (NT004 NN) JJ (NT004 NN) IN CD , PRP (NT006 (NT006 VBD (NT001 DT NN)) WDT (NT006 (NT006 VBD (NT005 VBN

(NT002 IN NN))) (NT011 CC (NT005 VBN (NT002 (NT002 IN (
NT001 DT JJ NN)) (NT002 IN (NT001 PRP$ NNS NN)))))))
.)

(((VBG IN (RB (NN JJ NN)) (IN (CD))) , ((PRP) (
VBD ((DT NN) ((WDT) (VBD ((VBN (IN (NN))) CC (VBN
(IN (DT JJ NN (IN (PRP$ NNS NN)))))))))))) .)

number = 17, pare1 = 8, pare2 = 10, cross = 6

((NT003 NNP NNP) (NT006 VBD (NT001 VBG JJ NNS)) , (NT005
JJ (NT002 IN NN)) CC (NT004 NN) (NT011 CC (NT008 NNS))
.)

(((NNP NNP) (VBD (VBG ((JJ NNS) , (JJ (IN (NN CC NN
CC NNS))))))) .)

number = 18, pare1 = 7, pare2 = 10, cross = 0

(PRP RB (NT007 VBD) CD (NT004 NN) WDT (NT006 (NT007 VBD)
(NT008 (NT004 NN NN) (NT008 NNS))) .)

(((PRP) RB (VBD ((CD NN) ((WDT) (VBD (NN NN NNS))
)))) .)

number = 19, pare1 = 18, pare2 = 22, cross = 44

((NT002 IN (NT001 DT JJ NN)) , PRP (NT006 (NT006 (NT006
(NT007 VBD VBG) RP (NT001 DT NN)) (NT011 CC (NT004 NN NN
) (NT002 (NT002 IN (NT001 DT NN NN NN) NN NNS) (NT002 IN
NN NN)))) CC (NT004 NN)) , (NT002 RB (NT002 (NT002 IN

NNS) (NT002 IN JJ NN))) .)

(((IN (DT JJ NN)) , ((PRP) (VBD (VBG RP (DT NN CC NN NN (IN (DT NN))) (NN (NN (NN ((NNS (IN (NN NN CC NN))) , (RB IN (NNS (IN (JJ NN)))))))))))))) .)

number = 20, pare1 = 7, pare2 = 11, cross = 5

((NT001 DT NN) (NT006 VBD (NT002 (NT002 (NT002 TO VB NN NNS) (NT002 IN NNS)) (NT002 IN JJ NN))) .)

(((DT NN) (VBD (TO (VB (NN NNS) (IN (NNS (IN (JJ NN))))))))) .)

B.3 方法 3

number = 1, pare1 = 18, pare2 = 28, cross = 2

((NT001 DT NN) POS (NT003 NN) , WDT (NT020 VBD) CD (NT011 NNS) IN JJS (NT020 VBD) PRP (NT001 DT NN) , (NT020 VBD) (NT004 NNP) (NT001 DT JJ NN) TO VB VBN (NT019 IN DT CD (NT003 NN)) VBG (NT018 (NT004 NNP) (NT012 (NT003 NN) (NT011 NNS))) TO JJS (NT012 JJ (NT011 NNS)) .)

((((DT NN) POS NN) , ((WDT) (VBD (CD NNS (IN ((JJS) (VBD (PRP) (DT NN))))))) ,) (VBD (NNP) ((DT JJ NN) (TO VB (VBN (IN ((DT CD NN) (VBG (NNP NN NNS) (TO ((JJS JJ) NNS))))))))) .)

number = 2, pare1 = 13, pare2 = 17, cross = 2

((NT001 DT NN) MD VB (NT022 JJR (NT003 NN)) (NT002 IN (

NT003 NN)) , (NT003 NN) CC (NT012 (NT003 NN) (NT011 NNS)) VBN IN (NT004 NNP) IN (NT004 NNP) POS (NT022 JJR (NT003 NN)) .)

((DT NN) MD (VB ((JJR) NN (IN (((NN) , (NN CC NN NNS)) (VBN (IN (NNP))))) (IN ((NNP) POS (JJR) NN)))) .)

number = 3, pare1 = 24, pare2 = 24, cross = 20

((NT004 NNP NNP NNP) (NT003 NN) (NT010 (NT005 NNS VBD) (NT011 NNS (NT002 (NT002 IN NN NNS) (NT002 IN (NT001 DT JJ NN))))) IN (NT004 NNP) TO VB (NT015 TO (NT011 NNS)) , (NT022 VBG (NT003 NN)) (NT002 IN (NT003 NN)) CC (NT012 (NT003 NN) (NT011 NNS)) IN JJ (NT003 NN) CC (NT012 JJ (NT012 (NT003 NN) (NT011 NNS))) .)

((NNP NNP NNP NN NNS) (VBD (NNS (IN (NN NNS (IN (DT JJ NN (IN (NNP))))))) (TO (VB (TO (NNS))) , (VBG (NN (IN ((NN) CC (NN NNS (IN ((JJ NN) CC (JJ NN NNS)))))))))) .)

number = 4, pare1 = 13, pare2 = 15, cross = 3

((NT011 NNS (NT002 IN NNS)) (NT016 (NT006 VBD (NT008 RB (NT002 IN (NT003 NN))))) , CC (NT003 NN) (NT010 (NT005 NNS VBD) (NT008 VBG NNS (NT002 IN (NT001 DT NN)))) .)

(((NNS (IN (NNS))) (VBD RB (IN (NN)))) , CC ((NN NNS) (VBD (VBG (NNS)) (IN (DT NN)))) .)

number = 5, pare1 = 17, pare2 = 20, cross = 16

(DT (NT004 NNP) (NT016 (NT006 VBD (NT002 (NT002 IN PRP$ JJ NN) (NT002 IN NNS)))) , (NT004 NNP NNP) , (NT016 (NT006 VBD (NT002 IN NNS))) IN JJ (NT003 NN) , (NT003 NN) CC (NT003 NN) TO (NT012 JJ (NT012 (NT003 NN) (NT011 NNS))) .)

((DT NNP) (VBD (IN (((PRP$ JJ NN (IN (NNS))) , (NNP NNP) ,) (VBD (IN (NNS (IN ((JJ NN) , (NN) CC (NN (TO (JJ NN NNS)))))))))))) .)

number = 6, pare1 = 3, pare2 = 8, cross = 0

(DT JJ (NT004 NNP) POS (NT003 NN) VBZ JJ (NT004 NNP) .)

((DT (((JJ) NNP) POS NN)) (VBZ (JJ (NNP))) .)

number = 7, pare1 = 13, pare2 = 20, cross = 0

(DT (NT004 NNP NNP) (NT003 NN) (NT016 (NT006 VBD (NT002 IN (NT001 DT NN)))) (NT019 IN JJ (NT004 NNP)) CC DT CD (NT011 NNS) (NT009 RB) , VBG IN JJS CD (NT004 NNP) , (NT016 (NT006 VBD VBN VBN)) .)

(((DT NNP NNP NN) (VBD (IN (DT NN)) (IN ((JJ) NNP)))) CC (((DT CD NNS RB) , (VBG ((IN (JJS)) (CD (NNP))))) ,) VBD (VBN (VBN))) .)

number = 8, pare1 = 10, pare2 = 12, cross = 9

((NT001 DT NN) (NT016 (NT006 VBD (NT008 RB (NT002 IN (NT001 DT JJ NN))))) (NT020 VBD) RP IN (NT004 NNP) , (NT009 RB) TO (NT004 NNP) .)

((DT NN) (VBD) (RB (IN ((DT JJ NN) (VBD RP (IN (NNP))) , RB (TO (NNP))))) .)

number = 9, pare1 = 17, pare2 = 26, cross = 10

((NT004 NNP NNP) (NT003 NN) VB (NT020 VBD) DT (NT004 NNP) (NT003 NN) IN PRP MD (NT009 RB) VB (NT015 TO VB (NT001 PRP$ NNS) (NT002 IN (NT003 NN))) CC JJ (NT003 NN (NT002 IN (NT001 DT NN))) (NT014 (NT007 PRP VBD) (NT002 IN CD NNS)) IN (NT004 NNP) .)

(((NNP NNP NN) (VB)) (VBD (DT NNP NN) (IN ((PRP) MD RB (VB (TO (VB (PRP$ NNS (IN ((NN) CC (JJ NN (IN (DT NN ((PRP) (VBD (IN (CD NNS)) (IN (NNP))))))))))))))))) .)

number = 10, pare1 = 6, pare2 = 10, cross = 3

((NT007 PRP VBD) (NT015 TO VB JJ) (NT003 NN) , VBG (NT001 PRP$ NN) (NT002 IN (NT003 NN)) .)

((PRP) (VBD (TO (VB (JJ NN)))) , (VBG (PRP$ NN (IN (NN)))) .)

number = 11, pare1 = 11, pare2 = 17, cross = 8

((NT004 NNP) (NT023 VBZ VBN) TO VB TO (NT004 NNP NNP) (NT003 NN) (NT015 TO VB (NT002 IN (NT001 DT JJ JJ NN))) POS (NT003 NN (NT002 IN CD NNS)) (NT002 IN (NT003 NN)) .)

((NNP) VBZ (VBN (TO (VB (TO (NNP NNP)) NN (TO (VB (IN ((DT (JJ JJ) NN) POS NN (IN (CD NNS (IN (NN)))))))))))) .)

number = 12, pare1 = 14, pare2 = 20, cross = 6

((NT001 DT NN) (NT023 VBZ VBN) RP (NT025 DT (NT011 NNS IN (NT004 NNP))) POS (NT004 NNP) (NT002 IN (NT001 DT JJ NN)) , WRB (NT010 (NT005 NNS (NT002 IN CD NNS) VBP) VBN (NT015 TO VB (NT012 JJ (NT011 NNS)))) .)

((DT NN) VBZ (VBN RP (DT NNS (IN ((NNP) POS (NNP))) (IN (DT JJ NN , ((WRB) ((NNS (IN (CD NNS))) VBP (VBN (TO (VB (JJ NNS))))))))))) .)

number = 13, pare1 = 13, pare2 = 20, cross = 0

(DT (NT004 NNP NNP) (NT003 NN) , (NT009 RB) , (NT020 VBD) TO (NT004 NNP) IN (NT004 NNP) TO VB VB (NT001 DT NN) IN (NT004 NNP) POS (NT011 NNS) (NT015 TO VB (NT001 DT JJ NN) (NT002 IN (NT003 NN))) .)

((DT NNP NNP NN) , RB , (VBD ((TO (NNP)) (IN (NNP))) (TO (VB (VB (DT NN (IN ((NNP) POS NNS)) (TO (VB (DT JJ NN (IN (NN))))))))))) .)

number = 14, pare1 = 8, pare2 = 9, cross = 4

((NT004 NNP) POS (NT003 NN) (NT016 (NT013 (NT006 VBD (NT001 DT JJ NN)) IN (NT004 NNP))) POS (NT004 NNP) .)

(((NNP) POS NN) (VBD (DT JJ NN) (IN ((NNP) POS (NNP))))) .)

number = 15, pare1 = 6, pare2 = 11, cross = 13

((NT004 NNP NNP NNP) (NT024 (NT016 (NT006 VBD (NT002 IN (NT001 DT NN)))) TO VB) PRP MD VB VBN .)

((NNP NNP NNP) (VBD (IN (((DT NN) (TO (VB (PRP)))) MD VB (VBN)))) .)

number = 16, pare1 = 16, pare2 = 14, cross = 18

(DT (NT004 NNP) , (NT009 RB) , (NT016 (NT006 (NT006 VBZ VBN (NT001 DT NN NNS)) (NT002 (NT002 IN (NT001 DT NN)) (NT002 IN (NT003 NN))))) MD VB (NT021 (NT008 VBN (NT002 IN VBG (NT001 DT NN)))) POS JJ (NT003 NN) .)

((DT NNP) , RB , VBZ (VBN (DT NN NNS) (IN ((DT NN (IN (NN))) MD VB (VBN (IN (VBG ((DT NN) POS JJ NN))))))) .)

number = 17, pare1 = 7, pare2 = 11, cross = 6

((NT001 DT JJ NN NN) (NT016 (NT006 VBD JJ)) (NT003 NN) , VBG PRP$ JJ (NT003 NN) IN (NT004 NNP) , MD VB VBN (NT009

RB) .)

((DT JJ NN NN) (VBD (((JJ NN) , (VBG (PRP$ JJ NN) (IN (NNP))) ,) MD VB (VBN RB))) .)

number = 18, pare1 = 8, pare2 = 9, cross = 5

((NT001 DT NN) , (NT021 (NT008 VBN (NT002 IN DT NNS))) POS JJ (NT003 NN) , (NT020 VBD) (NT009 RB) VB (NT001 DT NN) .)

(((DT NN) , (VBN (IN ((DT NNS) POS JJ NN))) ,) VBD RB (VB (DT NN)) .)

number = 19, pare1 = 7, pare2 = 8, cross = 2

((NT004 NNP NNP NNP) (NT016 (NT006 VBD JJ)) (NT003 NN) (NT004 NNP) CD IN VBN (NT012 JJ (NT011 NNS)) .)

((NNP NNP NNP) (VBD (JJ NN) (NNP CD) (IN ((VBN) JJ NNS))) .)

number = 20, pare1 = 17, pare2 = 21, cross = 6

((NT004 NNP NNP NNP) (NT010 (NT005 NNS VBD) (NT003 NN)) (NT004 NNP NNP) (NT002 IN (NT003 NN)) , VBG (NT004 NNP NNP) , WP (NT016 (NT013 (NT006 VBD (NT001 PRP$ NN)) IN (NT004 NNP)) (NT002 (NT002 IN CD NNS) (NT002 IN (NT001 DT NN)))) .)

((NNP NNP NNP NNS) (VBD ((NN) (NNP NNP)) (IN (NN))

) , (VBG ((NNP NNP) , ((WP) (VBD (PRP$ NN) (IN (NNP
)) (IN (CD NNS (IN (DT NN)))))))) .)