共起制約を組み込んだ確率文法による名詞句の統語的曖昧さの解消

田中, 省作
九州大学大学院システム情報科学研究科知能システム学専攻 : 博士後期課程

冨浦, 洋一
九州大学大学院システム情報科学研究科知能システム学専攻

日高, 達
九州大学大学院システム情報科学研究科知能システム学専攻

https://doi.org/10.15017/1513732

出版情報：九州大学大学院システム情報科学紀要. 5 (1), pp.69-74, 2000-03-24. 九州大学大学院システム情報科学研究院
バージョン：
権利関係：

九州大学大学院
システム情報科学研究科報告
第5巻　第1号　平成12年3月

Research Reports on Information Science and
Electrical Engineering of Kyushu University
Vol.5, No.1, March 2000

共起制約を組み込んだ確率文法による名詞句の統語的曖昧さの解消

田中省作* ・ 冨浦洋一**・ 日高 達**

Disambiguiation of Syntactic Structures of Japanese Noun Phrases "*NP* '*no*' *NP*" by PCFG Having Cooccurrence Constraints Embedded

Shosaku TANAKA, Yoichi TOMIURA and Toru HITAKA

(Received December 10, 1999)

Abstract: The noun phrase "*NP* '*no*' *NP*", that consists of two noun phrases connected by an adnominal particles '*no*', is frequently used in Japanese sentences. The surface structure of this phrase is simple, but it has various semantic structures. There has been proposed a grammar to get the semantic structure of this phrase systematically from its syntactic structure. And it is important to get a valid syntactic structure of an input noun phrase. However, there are several syntactic structures corresponding to the input noun phrase. This paper presents a Probabilistic Context Free Grammar(PCFG), which has coocurrence constraints embedded by subdividing a nonterminal with the semantic category of the headword of the phrase derived from it. Then the result of the experiment shows the effectiveness of this PCFG.

Keywords: Japanese noun phrase "*NP* '*no*' *NP*", Ambiguity of syntactic structure, Probabilistic context free grammar

1. はじめに

日本語文では, 二つの名詞句を助詞「の」で結合した名詞句「*NP* の *NP*」が頻繁に現れ, しかも, 多様な意味構造を持つことが知られている. このような名詞句「*NP* の *NP*」に対して, Montague の形式化に基づいて形式的に意味構造を与えるような文法体系が提案されている[8]. この文法体系では, 名詞句を意味的観点からさらに4つの統語範疇（普通名詞句 *CN*, 項句 *T*, 関係名詞句 *RN*, 事象名詞句 EN_k）に細分化することによって, 名詞句の統語構造と意味構造が対応付けている. 名詞句に対する統語規則は文脈自由文法のレベルで記述され, 各統語規則毎に意味構造を導出する規則（翻訳規則）が与えられている. つまり, 名詞句の統語構造を求めれば, 統語構造中の統語規則に対応する翻訳規則から一意に意味構造を導出することができる. よって, 名詞句の意味解析を統語解析のレベルで処理することができ, 統語解析時の統語的曖昧さの解消は, 極めて重要である. そこで, 本論文では, 文献 8) の文法に対して共起制約を組み込んだ文法を設計し, さらに文脈自由文法の統語規則に適用確率を付与した確率文脈自由文法(PCFG)を用いて, この名詞句の統語的曖昧さの解消を試みる.

まず, 2章で文献 8) の文法を概括し, 名詞句の統語的曖昧さの問題を示す. 3章では, 名詞句の統語的曖昧さを解消するために, この文献 8) の文法に共起制約を組み込む. 文献 8) の文法では単に統語範疇を表していた非終端記号を, それが導出する句の主辞の意味範疇で細分化することによって, 意味範疇間の共起制約を統語規則として表現する. さらに, 共起制約を組み込んだ文法を確率化した PCFG を, 名詞句の統語的曖昧さ解消に適用する. 4章では, 実際にコーパスより抽出した名詞句を用いた統語的曖昧さ解消の実験について述べる.

平成11年12月10日受付

* 知能システム学専攻博士後期課程
** 知能システム学専攻

2. 名詞句「*NP* の *NP*」の意味構造

本節では, 文献 8) の文法のうち本研究を述べる上で必要となる事柄のみを概観する. 詳細は文献 8) を参照して頂きたい.

2.1 名詞句の統語範疇

文献 8) の文法では, 名詞句「*NP* の *NP*」に対して, 統語規則と意味構造への翻訳規則とを対応づけるために, 従来, 単一の統語範疇として扱われていた名詞句を意味的観点から 4 つの統語範疇に細分化する.

2.1.1 普通名詞句（*CN*）

「色白の美人」における「色白」や「美人」は, それぞれ"色白である", "美人である"という性質を表している. このような名詞句の統語範疇は, 普通名詞句 (common noun phrase: *CN*) である. 統語的には, 「ある」「その」「すべての」などの英語の冠詞に相当する連体詞（本論文では, これらを限定詞 (determiner: *Det*) と呼ぶ）が結合し, 項句になるような句が *CN* である. また, この統語

範疇の名詞を, 特に普通名詞(common noun)とよぶ.

2.1.2 項句 (T)

ある特定の個体や事象を指示しているような名詞句の統語範疇は, 項句 (term phrase: T) である. 統語的には, 「太郎」や「彼」や CN に限定詞が結合した「その人」や「すべての男性」というような句が T である. この統語範疇の名詞句は, 動詞の格要素や後述する関係名詞句の補項になり得る句である. また, この統語範疇の名詞を, 特に項(term)とよぶ.

2.1.3 関係名詞句 (RN)

個体や事象間の関係を表す名詞句, すなわち, 「T の NP」が指示する個体と T が指示する個体の関係を表す名詞句 NP の統語範疇は, 関係名詞句 (relation noun phrase: RN) である. 例えば, 「兄」という名詞句は, RN である. このとき, T である「二郎」を「の」で結合して「二郎の兄」というような名詞句 を構成することによって, "二郎と兄弟関係にある個体で, かつ年上の個体" を指示することになる. この統語範疇の名詞を, 特に関係名詞(relation noun)とよぶ.

2.1.4 事象名詞句 (EN_k)

事象を表している名詞句の統語範疇は, 事象名詞句 (event noun phrase: EN) である. 「勉強」や「考え」といったサ変動詞の語幹や, 動詞が名詞に転化した名詞などは EN である. これらの EN は, 「太郎の英語の勉強」「太郎の考え」のように, 「の」を介して元の動詞に対応した格要素を取り, 事象を指示する. EN は取り得る格要素の数によって細分化され, k 個の格要素を取り得る事象名詞句の統語範疇を EN_k で表す. ただし, EN_0 は T である. この統語範疇の名詞を, 特に事象名詞(event noun)とよぶ.

2.2 名詞句の統語構造と意味構造

文献 8) の文法では, 「NP_1 の NP_2」の意味構造は, 細分化した名詞句の統語範疇と密接に対応している. その意味構造は, NP_1 が T か CN かで大別され, 以下のようになる.

- 「CN の NP」
  「CN の」は形容詞的に働き, 「CN の NP」の統語範疇は NP である. 意味的には, 「CN の」が NP の指示対象を制限する.
- 「T の NP」
  「T の」は, 「T が」や「T を」のような格要素のように働き, 「T の NP」の統語範疇は T となる. 意味的には, T と NP の間に意味関係†1が成立し, NP の特定の個体を指示することになる.

†1 NP が RN である場合は, RN の主辞の関係名詞, NP が EN_k である場合は, 主辞の事象名詞との格関係が意味関係に相当する. NP が CN または T である場合は, 『所有関係』や『位置関係』といった意味関係を T,NP に応じて補う[6),7)].

Table-1 Syntactic rules in the proposed grammar[8)].

$CN \longrightarrow n$	(n is a common noun)	(1)
$T \longrightarrow n$	(n is a term)	(2)
$RN \longrightarrow n$	(n is a relation noun)	(3)
$EN_k \longrightarrow n$	(n is an event noun)	(4)
$T \longrightarrow CN$		(5)
$T \longrightarrow RN$		(6)
$EN_{k-1} \longrightarrow EN_k$		(7)
$T \longrightarrow Det\ CN$		(8)
$T \longrightarrow T$の RN		(9)
$EN_{k-1} \longrightarrow T$の EN_k		(10)
$T \longrightarrow T$の CN		(11)
$T \longrightarrow T$の T		(12)
$CN \longrightarrow CN$の CN		(13)
$RN \longrightarrow CN$の RN		(14)
$EN_k \longrightarrow CN$の EN_k		(15)
$T \longrightarrow CN$の T		(16)

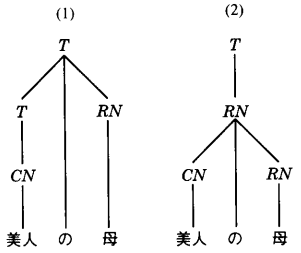

Fig.1 Syntactic structures of "*bijin no haha*"((1)a mother of a beauty or (2)a beautiful mother).

文献 8) で列挙されてある名詞句に対する統語規則を **Table–1**に示す. 文献 8) の文法では, 統語規則は文脈自由文法(context free grammar: CFG)で記述され, 各統語規則毎に翻訳規則が与えられている. よって, 名詞句「NP の NP」の意味の相違は, 意味構造の曖昧さ(意味的曖昧さ)として, 統語構造における曖昧さ(統語的曖昧さ)に反映される. 例えば, 「美人の母」という場合は,

(1) "美人を娘にもつ母"
(2) "美人である母"

という 2 通りの解釈ができ, それぞれの意味に対応する統語構造は**Fig.1**である. (1)の解釈では, 「美人の」の「美人」はある特定の個体を指示しており, その統語範疇は T である. つまり, CN の「美人」に付加される「その」や「ある」といった限定詞が省略されたことで, 表層的には変化なく, T に範疇変換していると考えるられる. 一方, (2)の解釈では, 「美人の」は形容詞的に機能し, 「母」が指示する対象を限定しているので, 「美人」の統語範疇は CN である.

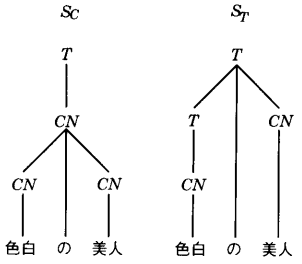

Fig.2 Syntactic structures of "*irojiro no bijin*"(a fair beauty).

2.3 名詞句の統語的曖昧さ

文献 8) の文法は, 単に名詞句の統語構造と意味構造の関係を示したもので, そのまま統語解析に適用すると次のような統語的曖昧さが生じる. 例えば, 「色白の美人」という名詞句は, 統語解析すると**Fig.2**の 2 つの統語構造 S_C, S_T が得られる. この場合, 意味を考えれば, 「色白の美人」の統語構造としては S_C が妥当で, S_T は妥当性を欠く. また, 「色白の NP」という名詞句が常に S_C のような統語構造を構成するわけではなく, 共起する NP によって, その統語構造が決まる. 例えば, 「色白の原因」という名詞句では, 「色白の」は S_T のように「T の」として NP に係る統語構造となる. つまり, **名詞句「NP_1 の NP_2」において NP_1, NP_2 がどの統語範疇で「の」と結合し共起するかということは, NP_1, NP_2 が導出する句に強く依存する**.

本研究では, NP_1, NP_2 が導出する句の組み合わせは膨大となるので, それぞれの句の主辞だけを考慮することにする. そこで, 名詞句を解析する文法には NP_1, NP_2 が導出する句の主辞を表現する必要がある. 本研究では, 文献 8) の文法に NP_1, NP_2 の主辞の意味範疇間に共起制約を組み込むことを考える. CFG に共起制約を組み込む手法として, 文献 3),5) が既に提案されており, これらは, 非終端記号を細分化することで意味範疇間の共起制約を統語規則として表現する†2. 本研究では, 文献 5) の手法に従い, 文献 8) の文法に対して共起制約を組み込んだ文法を設計する.

3. 意味範疇間の共起制約を組み込んだ PCFG の設計

3.1 非終端記号の意味範疇による細分化と意味範疇間の共起制約の組み込み

本節では, 文献 8) の文法において統語範疇そのものを表していた非終端記号を, その非終端記号が導出する句の主辞の意味範疇によって細分化することによって, 意味範疇間の共起制約を統語規則で表現する. まず, 記法の定義を行う.

定義 1 ($\widetilde{w}$)
単語 w の意味範疇の1つを $\widetilde{w}$ で表す†3. □

文献 8) の文法では, 非終端記号は統語範疇と等価であったが, その非終端記号を統語範疇と主辞の意味範疇の組で表す. これは, 非終端記号をそれが導出する主辞の意味範疇によって細分化することに相当する. 非終端記号の記法について定義しておく.

定義 2 (非終端記号 $X(\widetilde{\alpha})$)
$X(\widetilde{\alpha})$ は, 統語範疇が X で, 主辞の意味範疇が $\widetilde{\alpha}$ の句を導出する非終端記号である. □

このような形で非終端記号を細分化することによって, 文献 8) の文法の統語規則(1)~統語規則(16)は, 次のようになる.

統語規則の右辺に終端記号または非終端記号が1つ現れる統語規則(1)~統語規則(7)は,

$$X(\widetilde{n}) \longrightarrow n \tag{17}$$
$$X(\widetilde{n}) \longrightarrow Y(\widetilde{n}) \tag{18}$$

となる. ただし, $X, Y \in \{T, CN, RN, EN_k\}$ で, n は名詞である. 統語規則(18)は, **主辞の意味範疇が $\widetilde{n}$ である統語範疇 Y の句が, 限定詞が省略されて統語範疇 X の句に範疇変換する**ということを表している. 次に, 名詞句に限定詞が付加する統語規則(8)については,

$$T(\widetilde{n}) \longrightarrow Det\ CN(\widetilde{n}) \tag{19}$$

となる. 統語規則の右辺に名詞句が2つ現れる統語規則(9)~統語規則(16)は,

$$X(\widetilde{n_2}) \longrightarrow Y(\widetilde{n_1})\ の\ Z(\widetilde{n_2}) \tag{20}$$

となる. ただし, $X, Y, Z \in \{CN, T, RN, EN_k\}$ である. この統語規則(20)は, **主辞の意味範疇が $\widetilde{n_1}$ である統語範疇 Y の句が「の」と結合し, 主辞の意味範疇が $\widetilde{n_2}$ である統語範疇 Z の句と共起し得, その結果, X に範疇変換する**ということを表している.

文献 8) の文法にこのような方法で, 共起制約を組み込んだ文法は, **Table–2**のようになる. また, 各統語規則に対応する翻訳規則も, 元の文献 8) の文法の統語規則と翻訳規則の対応関係を基に機械的に記述され, 統語規則と翻訳規則の厳密な対応関係も保存される.

この文法で, 前の例で挙げた「色白の美人」を解析すると, 「色白」が統語範疇 T で「の」と結合して, CN の「美人」と共起するような統語規則 $T(\widetilde{美人}) \longrightarrow T(\widetilde{色白})\ の\ CN(\widetilde{美人})$ は共起制約を満たさ

†2 正確には, 係り受け制約を組み込んだもので, 文献 5) では係り受け制約を 係る語・係られる語・係りの種類 の 3 つで記述している.

†3 後述の実験では, w が多義語の場合, w が出現した単文を参照し, 人手で $\widetilde{w}$ を決定した.

Table-2 Syntactic rules in the grammar having cooccurence constraints embedded.

$CN(\widetilde{n}) \longrightarrow n$　(n is a common noun)
$T(\widetilde{n}) \longrightarrow n$　(n is a term)
$RN(\widetilde{n}) \longrightarrow n$　(n is a relation noun)
$EN_k(\widetilde{n}) \longrightarrow n$　(n is an event noun)
$T(\widetilde{n}) \longrightarrow CN(\widetilde{n})$
$T(\widetilde{n}) \longrightarrow RN(\widetilde{n})$
$EN_{k-1}(\widetilde{n}) \longrightarrow EN_k(\widetilde{n})$
$T(\widetilde{n}) \longrightarrow Det\ CN(\widetilde{n})$
$T(\widetilde{n_2}) \longrightarrow T(\widetilde{n_1})$ の $RN(\widetilde{n_2})$
$EN_{k-1}(\widetilde{n_2}) \longrightarrow T(\widetilde{n_1})$ の $EN_k(\widetilde{n_2})$
$T(\widetilde{n_2}) \longrightarrow T(\widetilde{n_1})$ の $CN(\widetilde{n_2})$
$T(\widetilde{n_2}) \longrightarrow T(\widetilde{n_1})$ の $T(\widetilde{n_2})$
$CN(\widetilde{n_2}) \longrightarrow CN(\widetilde{n_1})$ の $CN(\widetilde{n_2})$
$RN(\widetilde{n_2}) \longrightarrow CN(\widetilde{n_1})$ の $RN(\widetilde{n_2})$
$EN_k(\widetilde{n_2}) \longrightarrow CN(\widetilde{n_1})$ の $EN_k(\widetilde{n_2})$
$T(\widetilde{n_2}) \longrightarrow CN(\widetilde{n_1})$ の $T(\widetilde{n_2})$

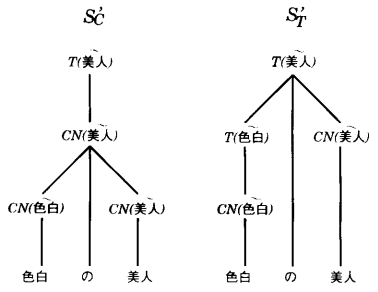

Fig.3 Syntactic structures of "*irojiro no bijin*" on the grammar having cooccurrence constraints embedded.

ないため獲得されない．その結果，「色白の美人」に対する統語構造は，S'_C のみが得られることになる(**Fig.3**)．

しかし，このように意味範疇間の共起制約を組み込んだ文法でも，例えば，「美人の母」に対しては，**Fig.4** のような統語構造が得られ，依然，統語的曖昧さは解消されない場合がある．そこで，意味範疇間の共起制約を組み込んだ文法をさらに確率化し，確率文脈自由文法(probabilistic context free grammar: PCFG)とする[2]．PCFG は，CFG の各統語規則に適用確率を付与したもので，統語構造に対して生起確率を計算することができる．そして，付与された生起確率に基づいて入力の可能な統語構造間に優先順位を与えることができる．

3.2 確率文脈自由文法

定義 3 (確率文脈自由文法)
PCFG G は，$G = \langle N, \Sigma, P, S, p \rangle$ の5字組で定義される．ただし，それぞれの記号の意味は以下の通りである．

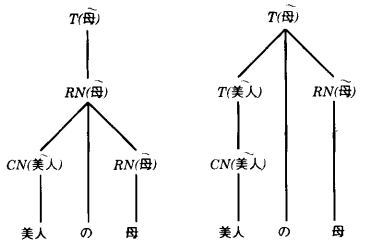

Fig.4 Syntactic structures of "*bijin no haha*" on the grammar having cooccurrence constraints embedded.

N：非終端記号の有限集合
Σ：終端記号の有限集合
P：$\{X \longrightarrow \alpha \mid X \in N, \alpha \in (\Sigma \cup N)^*\}$ の有限部分集合
S：開始記号 ($S \in N$)
p：$P \rightarrow (0,1]$ (P から区間 $(0,1]$ への写像)

統語規則 $X \longrightarrow \alpha_i$ $(i = 1, 2, \cdots, I_X)$ の適用確率は，$p(X \longrightarrow \alpha_i)$ で表される．統語規則の左辺が同一の非終端記号である適用確率の総和については次式が成り立つ．

$$\sum_{i=1}^{I_X} p(X \longrightarrow \alpha_i) = 1$$

□

PCFG で統語構造 T に対する生起確率 $P(T)$ は，T の導出に適用された統語規則の適用確率の積 $\prod_{\delta \in P} p(\delta)^{n(T,\delta)}$ となる．但し，$n(T,\delta)$ は，統語構造 T の導出で統語規則 $\delta \in P$ が適用された回数を表す．

次に，PCFGの確率パラメタ p の推定について述べる．確率パラメタは，通常，予め収集された統語構造列 $\langle T_1, T_2, \cdots, T_N \rangle$ をサンプルとして統計的に推定される．本論文では，尤度（サンプルの発生確率）

$$L(p\,;\langle T_1, \cdots, T_N \rangle) = \prod_{i=1}^{N} P(T_i; p)$$

が最大になるように確率パラメタ p を推定する最尤推定法を用いる．

定理 1 (最尤推定法)
尤度 $L(p; \langle T_1, \cdots, T_N \rangle)$ を最大にする確率パラメタ p の推定値 $\hat{p}$ は次式で与えられる．

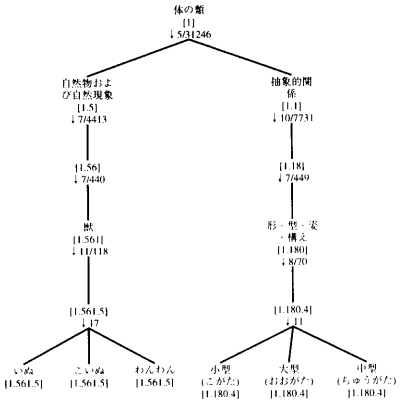

Fig.5 A part of *Bunruigoihyou*.

$$\hat{p}(X \longrightarrow \alpha_i) = \frac{\sum_{k=1}^{N} n(T_k, X \longrightarrow \alpha_i)}{\sum_{i=1}^{I_X} \sum_{k=1}^{N} n(T_i, X \longrightarrow \alpha_i)} \tag{21}$$

□

共起制約を組み込んだ文法を確率化した PCFG では, 意味範疇間の共起制約を考慮した形で, 統語構造間に優先順位が付与することができる.

4. 実　　験

本節では, 文献 8) の文法を確率化した PCFG と, 共起制約を組み込んだ文法を確率化した PCFG を用いた名詞句の統語的曖昧さ解消に関する実験について述べる.

4.1 実験データ

実験データとして, EDR コーパス[1]より抽出し, それぞれの文法にあわせて統語構造に直した 8,738個の名詞句「*NP* の *NP*」を用いた. この統語構造列を $\mathcal{S}$ と記す.

4.2 データ・スパースネスへの対処

本実験では, 各単語の意味範疇は, 基本的には分類単語表[4]上の各単語の語義にあたる項目に対応させる(**Fig.5**). 例えば, 統語構造が「*CN* の *CN*」の「大型の犬」を構成する統語規則は,

$$CN(\langle\langle \text{犬} \rangle\rangle) \longrightarrow CN(\langle\langle \text{大型} \rangle\rangle) \text{ の } CN(\langle\langle \text{犬} \rangle\rangle) \tag{22}$$

となる.

しかし, 共起制約を組み込んだ文法で意味範疇を単語の語義のように細かく設定すると, 統語規則の数が膨大となり, 信頼性の高い推定を行えるだけの十分に大きなサンプルが得られないというデータ・スパースネスの問題が生じる. このデータ・スパースネスに対処するため, 本実験では, 各単語の意味範疇を, 分類単語表において語義にあたる項目から, さらに上位の項目に汎化したものも考える. 例えば, 意味範疇を単語の語義にあたる項目から 1 段汎化したものを考えた場合, 統語規則(22)は, 分類語彙表において《犬》,《大型》の直接の上位概念がそれぞれ《獣》,《形・型・姿・構え》であるので,

$$CN(\langle\langle \text{獣} \rangle\rangle) \longrightarrow CN(\langle\langle \text{形・型・姿・構え} \rangle\rangle) \text{ の } CN(\langle\langle \text{獣} \rangle\rangle)$$

となり,「小柄の熊」などもこの統語規則によって解析できるようになる. ただし, $\langle\langle\alpha\rangle\rangle$ は α という意味範疇を表す.

4.3 実験手順

実験では, 次のような PCFG を用いて, その解析結果を比較する.

- G: 文献 8) の文法を確率化した PCFG
- G'_ℓ: 共起制約を組み込んだ文法を確率化した PCFG ただし, ℓ は, 非終端記号中の意味範疇を, 分類単語表上で汎化した段数を表す. 本実験では, $\ell = 0, 1, 2$ を考えた.

本実験では, 次の2通りを行う.

1. **クローズド・データ実験**: 学習データをそのまま評価データとする.
2. **オープン・データ実験**: 学習データと評価データを別のものとする. 本実験の場合, データ数が十分とは言えないので, クロス・ヴァリデーションによって評価を行う. 実験の手順は以下の通り.

   (Step 1) サンプルを $\mathcal{S} = \mathcal{S}^1 \cup \mathcal{S}^2 \cup \cdots \cup \mathcal{S}^{10}$ をランダムに10分割する. ただし, $\mathcal{S}^i \cap \mathcal{S}^j = \emptyset$ $(i \neq j)$.

   (Step 2) 分割した $\mathcal{S}^1 \sim \mathcal{S}^{10}$ のうちの1つ $\mathcal{S}^i$ を評価データとし, 残りの $\bigcup_{\substack{1 \leq j \leq 10 \\ j \neq i}} \mathcal{S}^j$ を文法の学習データとして用いる. このようにして構成した PCFG で評価データ $\mathcal{S}^i$ を統語解析し, 後述の評価値(解析可能率, 適合率, 再現率)を計算する.

   (Step 3) (Step 2)の i を1から10まで変え, 各 i の評価値(解析可能率, 適合率, 再現率)の平均を本実験における評価値(解析可能率, 適合率, 再現率)とする.

4.4 実験結果

実験結果として, 解析可能率, 適合率, 再現率を示す. まず, 解析可能率とは, 統語解析できた名詞句の割合である. 次式で計算される.

Table-3 The result of the experiment of grammar G.

closed data	open data		
Rec.	Ana.	Prec.	Rec.
93.0%	100%	93.0%	93.0%

'Ana.' is the rate of analyzable sentences in input sentences. 'Prec.' is the precision. 'Rec.' is the recall, that is the product of 'Ana.' and 'Prec.'.

Table-4 The result of the experiment of grammar G'_ℓ.

ℓ	closed data	open data		
	Rec.	Ana.	Prec.	Rec.
0	99.9%	10.5%	99.6%	10.45%
1	99.8%	12.9%	99.3%	12.80%
2	99.7%	20.0%	98.4%	19.68%

$$\text{解析可能率} = \frac{\text{統語解析できた名詞句の数}}{\text{名詞句の数}}$$

適合率は, 統語解析できた名詞句のうち, 正しい統語構造が最も高い生起確率であった名詞句の割合である. 次式で計算される.

$$\text{適合率} = \frac{\text{正しい統語構造が得られた名詞句の数}}{\text{統語解析できた名詞句の数}}$$

再現率は, 統語解析でき, なおかつ正しい統語構造が最も高い生起確率であった名詞句の割合である. よって, 再現率は解析可能率と適合率の積で求められる.

文法 G の実験結果を**Table–3**に示す. ただし, 'Ana.' は解析可能率, 'Prec.' は適合率, 'Rec.' は再現率を表す. 文法 G'_ℓ の実験結果を**Table–4**に示す.

4.5 考察

クローズド・データ実験では, 共起制約を組み込み確率化した文法 G'_ℓ が, 文献 8) の文法を確率化した文法 G に比べ, 非常に高い正解率で統語構造を決定することができた.

また, オープン・データ実験では, G'_ℓ が解析できた名詞句に対しては, G に比べやはり非常に高い正解率で統語構造を決定できることが示された. しかしながら, 解析可能率の低さが大きな問題である. 単純には, サンプルとして十分なデータ数が獲得されれば, G'_ℓ の正解率は, クローズド・データ実験における G'_ℓ の正解率に近づくと予想される. しかし, 解析に失敗する多くの原因は,

$$X(\widetilde{n_2}) \longrightarrow Y(\widetilde{n_1}) \text{ の } Z(\widetilde{n_2})$$

という意味範疇間の共起制約を表す統語規則の洩れである. そこで, この統語規則の単純な組み合わせの数を考えてみると, 分類語彙表に登録されている名詞の語義が約3,500程度であることから, $\ell = 0$ では 10^6 の定数倍程度となる. $\ell = 1$ でも 10^5 の定数倍程度であり, 実際には十分な大きさのサンプルを収集することは難しい. パラメタ数は, $\ell = 2$ のとき 10^4 の定数倍程度となり, 現実的にはこの程度に抑えなければならないと思われる.

また, サンプルを大きくしていくと同時に, このようなデータ・スパースネスに対する対処法として既に提案されている手法を適用することも必要である.

5. ま　と　め

本研究では, 名詞句「*NP* の *NP*」の統語的曖昧さに対して, 文献 8) の文法を基に意味範疇間の共起制約を組み込んだ PCFG を設計し, その有効性を確認した. 今後は, データ・スパースネスへの対処が重要である. 文献 9) にあるシソーラス上の上位-下位関係を完全に文法に取り込み解析可能率を向上させる手法を適用することを検討している.

参　考　文　献

1) 日本電子化辞書研究所: EDR 電子化辞書仕様説明書 (1995).
2) 日高 達: 確率文法, 情報処理学会学会誌, Vol. 36, No. 2, pp. 169–176 (1995).
3) Hogenhout, W. R. and Matsumoto, Y.: Training Stochastical Grammars on Semantic Categories, *IJCAI'95 Workshop on New Approches to Learning for Natural Language Processing*, pp. 65–70 (1995).
4) 国立国語研究所: 分類語彙表, 秀英出版 (1964).
5) 田辺利文, 冨浦洋一, 日高 達: 係り受け文脈自由文法とその日本語への適用, 情報処理学会論文誌, Vol. 41, No. 1 (掲載予定).
6) 田中省作, 冨浦洋一, 日高 達: 統計情報を用いた名詞句「*NP* の *NP*」の意味関係の抽出法, 電子情報通信学会技術研究報告 NLC98-4 (1998).
7) 田中省作, 柳瀬康雄, 冨浦洋一, 日高 達: k-NN 推定法に基づいた名詞句の意味関係の推定, 九州大学大学院システム情報科学研究科報告, Vol. 4, No. 2, pp. 159–164 (1999).
8) 冨浦洋一, 中村貞吾, 日高 達: 名詞句「*NP* の *NP*」の意味構造, 情報処理学会論文誌, Vol. 36, No. 6 (1995).
9) トウシンバット, D., 冨浦洋一, 日高 達: 係り受け文脈自由文法の強化法, 情報処理学会研究会報告 NL128-6 (1998).