社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
IT2002-30 (2002-09)

計算論的情報源モデルと圧縮アルゴリズム

中澤　真†　松嶋　敏泰††　平澤　茂一††

† 早稲田大学メディアネットワークセンター
〒169-8050 東京都新宿区戸塚町 1-104
†† 早稲田大学理工学部経営システム工学科
〒169-8555 新宿区大久保 3-4-1
E-mail: †nakazawa@mn.waseda.ac.jp

あらまし　情報源符号化における無歪み圧縮の手法として，形式文法を用いた圧縮法の研究が行われている．文法に基づく符号化法は，サンプルデータから計算機モデルを推定する問題とも考えることができる．これまでの研究では，情報源系列から符号化される文法は文脈自由文法を用いていたが，これが最適というわけではない．文法のクラスを変更することにより，冗長度が0に収束する速度や，符号化・復号化の計算量をより適切なものにする可能性があるためである．そのため，情報源クラスに応じた文法のクラスを考える必要がある．本稿では，情報源系列の発生過程も計算機モデルとする計算論的な情報源を考え，その性質と形式文法の階層構造がどのような影響を与えるのか明らかにする．また，情報源が既知の場合における符号化アルゴリズムを示し，このときの計算量を明らかにする．最後にこれらの情報源におけるユニバーサル符号を示す．
キーワード　確率的形式文法，無歪み圧縮，ユニバーサル符号，チョムスキーの階層

Computational Source Information Model and its Algorithm

Makoto NAKAZAWA†, Toshiyasu MATSUSHIMA††, and Shigeichi HIRASAWA††

† Media Network Center, Waseda University
Totsuka-cho 1-104, Shinjuku-ku, Tokyo, 169-8050 Japan
†† School of Science and Engineering, Waseda University
Ohkubo 3-4-1, Shinjuku-ku, Tokyo, 169-8555 Japan
E-mail: †nakazawa@mn.waseda.ac.jp

Abstract　Recently, grammar based codes are researched in the area of lossless source coding. However, it is not clear which formal grammar is appropriate in Chomsky hierarchy for information compression. In order to clarify this, we propose a computational information source model. Computational information source models generate messages in terms of their probabilistic formal grammars. We present each characterization for probabilistic regular grammar based information source and probabilistic context-free grammar based information source. In the case when the parameters of these information sources are known, the lossless compression algorithms and their time complexity are shown. At the last, it is shown that a grammar based code is a universal code with respect to the family of probabilistic context-free grammars information sources.
Key words　probabilistic formal grammars, lossless compression, universal coding, Chomsky hierarchy

1. はじめに

情報源符号化における無歪み圧縮の手法として，形式文法を用いた圧縮法の研究が行われている．最近の研究としてはSEQUITUR [9] [10] や MPM [3] [4] などいくつかの研究があるが，いずれも情報源系列を言語として受理するような制限付きの文脈自由文法を構成することにより圧縮を行う．形式文法は Chomsky の階層 [16] とよばれる階層構造を持ち，この階層はそれぞれ対応する計算機モデルが存在する．例えば，文脈自由文法はプッシュダウンオートマトン（PNFA）と等価であり，また最も外の階層に位置する句構造文法はチューリングマシンに相当する．しかし，文法に基づく符号において文脈自由文法

を用いることが最適というわけではない．文法のクラスを変更することにより，冗長度が0に収束する速度や，符号化・復号化の計算量をより適切なものにする可能性があるためである．そのため，情報源クラスに応じて文法のクラスを考える必要がある．

この問題を考えるため，情報源系列の発生過程も計算機モデルとして考えた計算論的情報源を定義する．これは情報源系列の出現過程を確率的形式文法に従って生成されるような情報源モデルであり，形式文法のクラスに応じて計算量による階層的な情報源クラスが構成されることになる．確率的形式文法は導出時にどの生成規則が適用されるか確率的に決定されるように，通常の形式文法を拡張した概念である．

本稿ではこの情報源についての性質を示し，形式文法の階層構造が情報源モデルにどのような影響を与えるのかを明らかにする．特に，確率的形式文法では非確率的な形式文法の性質が保存されるとは限らないため，確率的正則文法と確率的文脈自由文法を確率過程とする情報源に対する性質を示し，この情報源が既知の場合における符号化アルゴリズムとその計算量を示す．

最後に，Kieffer ら [3] がある文法に基づく符号が有限状態情報源についてユニバーサルであることを示した結果について，これらを計算論的情報源上でも成立することを示し，より一般的な情報源においてユニバーサル性が示せることを証明する．

2. 準　　備

まず初めに確率的文法を定義するための準備をする．有限かつ非空なアルファベットを Σ，Σ 上のすべての語(注1)の集合を Σ^* で表す．この Σ 上の語の順序対の有限集合を R とする．R の要素 (μ,ω), $\mu,\omega\in\Sigma^*$ を生成規則と呼び，$\mu\to\omega$ と表す．

生成規則の集合 R に対し，$\mu\in\Sigma^*$ を書き換えるすべての生成規則の集合を R_μ とする．このとき μ に対してどの生成規則が適用されるか確率的に決定されるとし，その確率を $P(r)$ で表す．確率の公理から $\sum_{r\in R_\mu}P(r)=1$ である(注2)．このとき確率的文法 G を以下のように定義する．

[定義 1]　Σ_N を非終端アルファベット，Σ_T を終端アルファベットとし，$\Sigma=\Sigma_N\cup\Sigma_T$，$\Sigma_N\cap\Sigma_T=\phi$ を満足する集合とする．このとき文法 G を $G=(R,\Sigma_N,\Sigma_T,S,P)$ と定義する．ただし，S は開始記号で $S\in\Sigma_N$ である．

確率的文法は言語の要素が確率的に生成されるモデルである．確率的文法が与えられたとき，言語とその要素である語の導出確率がどのように定まるかをこれから定義する．便宜上，非終端アルファベット Σ_N の要素を大文字のアルファベット $A,B,C,\ldots$ で，終端記号 Σ_T の要素を小文字のアルファベット $a,b.c,\ldots$ で表すこととする．

$\alpha,\beta,\nu_1,\nu_2\in\Sigma^*$ に対して，R の要素である生成規則 $\mu\to\omega$ について $\alpha=\nu_1\mu\nu_2$, $\beta=\nu_1\omega\nu_2$ が成り立つとき $\alpha\Rightarrow\beta$ と表し，文法 G において α から β が直接導出されるという．また，Σ 上の語の有限列 $\omega_1,\omega_2,\ldots,\omega_k$ $(k\geq 0)$ について $\alpha=\omega_1\Rightarrow\omega_2$, $\omega_2\Rightarrow\omega_3,\ldots,\omega_{k-1}\Rightarrow\omega_k=\beta$ が成り立つとき $\alpha\Rightarrow^*\beta$ と表し，文法 G において α から β が導出されるという．このことから導出は直接導出に対応する生成規則の列であり，導出 d は $d=(r_1,r_2,\ldots),r_i\in R$ となる．

[定義 2]　文法 G によって生成される言語 $L(G)$ を

$$L(G)\triangleq\{\omega\in\Sigma_T^*:S\Rightarrow^*\omega\}\qquad(1)$$

と定義する．

確率的文法では生成規則に割り当てられた確率からどの導出あるいは導出木が選択されるか確率的に定まる．またこれらの確率がきまれば産物である語の確率も定義できる．

[定義 3]　確率的文法 G の中のある導出 d に対し，その導出が適用される確率は $p_d(d)$ は

$$p_d(d)\triangleq\prod_{r\in d}P(r)\qquad(2)$$

となる(注3)．このとき文法 G によって語 ω が導出される確率 $p_\omega(\omega)$ は

$$p_\omega(\omega)\triangleq\sum_{d'\in\{d|Gen(d)=\omega\}}p_d(d')\qquad(3)$$

と定義する．ただし，$Gen(d)$ は導出 d によって生成された語を表す．

この確率的文法を確率モデルとする情報源を計算論的情報源とよぶことにする．この場合，確率的文法の確率に従い一つの導出木(注4)が生成され，これに従って導出された終端記号列が情報源系列となる．ただし，扱いを簡単にするため確率的文法は既約であり，かつ $\epsilon-$ 規則を含まないと仮定する．

[例 1]　確率的文法に基づく情報源の例として確率的正則文法の場合で説明する．$\Sigma_T=\{0,1\}$ とし，文法 G の生成規則集合 R が以下の要素からなるとする．生成規則の前の数値がその生成規則の確率を表す．

$$1:S\to 0A$$
$$0.3:A\to 10A$$
$$0.7:A\to 1$$

この言語 $L(G)$ は正規表現 $0(10)^*1$ を受理する言語であるが，$L(G)$ に含まれる個々の語は確率的に生起し，その確率は表1のようになる．

表1　確率的文法による語の生起確率

語	01	0101	010101	$\cdots$
生起確率	0.7	0.21	0.063	$\cdots$

(注1)：0個以上の Σ 上の要素（記号）からなる有限列．

(注2)：確率的句構造文法，確率的文脈依存文法を考える場合には若干の制約条件が必要となるが，一般性を失うことはない．

(注3)：$p_d(d)$ は文法 G に依存するため，$p_d(d|G)$ と表記すべきであるが文法 G が明らかな場合は簡略化して表記する．これは他の確率についても同様である．

(注4)：導出木と導出は一般には1対1対応しないが，導出を最左導出に限定すれば1対1の対応がつく．

表 2　Chomsky による言語階層

生成文法	言語型	言語名
type 0	L_0	句構造言語
type 1	L_1	文脈依存言語
type 2	L_2	文脈自由言語
type 3	L_3	正則言語

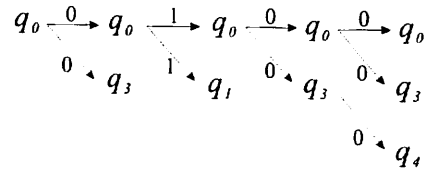

図 1　NFA の状態の増殖

計算論的情報源ではその形式文法の表現力に応じて情報源のクラスが定まる．ここでは形式文法のクラスとして Chomsky の階層 [16] に従った定義を以下に示す．

[定義 4]　文法 G の生成規則集合 P の各要素が次の条件 $0, 1, \cdots, i$, を満たすとき，type i 文法，$i = 0, 1, 2, 3$ という．

(条件 0)　$\mu \to \omega,\ \mu, \omega \in \Sigma^*$

(条件 1)　$\mu \to \omega,\ |\mu| \leq |\omega|,\ \mu, \omega \in \Sigma^*$ (注5)

(条件 2)　$A \to \omega,\ A \in \Sigma_N, \omega \in \Sigma^*$

(条件 3)　$A \to \alpha B$ または $A \to \alpha$

ただし，$A, B \in \Sigma_N, \alpha \in \Sigma_T^*$

正則文法は正則言語を受理する Chomsky の階層における最も制約が厳しい文法であり，また最も単純な文法である．

確率的文法のエントロピーとしては，Soule [12] によって導出エントロピーと文形式エントロピーの二つが定義されている．

[定義 5]　$\Omega(G)$ を開始記号 S で始まるすべての導出の集合とする．ある導出 $d \in \Omega(G)$ に対し，この導出 d の確率 $p_d(d)$ は導出において用いられたすべての生成規則の確率の積とする．このとき，導出エントロピー H_d を以下のように定義する．

$$H_d(G) = \sum_{d \in \Omega(G)} p_d(d) \log p_d(d) \qquad (4)$$

これに対し，文形式エントロピーは

$$H_s(G) = \sum_{\omega \in L(G)} p_\omega(\omega) \log p_\omega(\omega) \qquad (5)$$

3.　確率的正則文法に基づく情報源

計算論的情報源として，確率的正則文法に基づく情報源から考えてみる．正則文法の場合，左線形性あるいは右線形性のいずれかの性質をもつが，ここでは左線形性が成り立っていると仮定する(注6)．この情報源に対する符号化を考える前に，この情報源についての性質を示しておく．そのための準備として文法の曖昧性について定義する．

[定義 6]　文法 G に対し，その言語 $L(G)$ のある語 ω が 2 つの異なる導出木をもつとき文法 G は曖昧であるという．また言語 L を表現するどんな文法も曖昧であるとき，その言語は本質的に曖昧であるという．

文法が曖昧な場合には構文解析の手間が無曖昧な文法に比べて煩雑になる．そのため文法が受理する言語が本質的曖昧であるかどうかの決定は重要な問題である．

[定理 1]　正則言語（正則集合）は本質的曖昧ではない．

[証明]　正則言語はその定義から，これを受理する決定性有限オートマトン（DFA）が必ず存在する．DFA は同じ入力記号に対し，異なる状態に遷移することはないため，同じ入力に対して異なる導出で受理されることはありえない．このためどのような正則言語についても，これを導出する無曖昧な正則文法が必ず存在する．　□

定理 1 から，確率を含まない通常の正則文法では無曖昧な文法のみを考えれば十分であるが，確率という要素が加わった確率的文法では，本質的曖昧性の議論のみでは不十分な場合が存在することを次に示す．まず通常の形式言語における等価性の定義を確率的言語のために拡張する．

[定義 7]　確率的文法 G と G' によってそれぞれ定まる確率的言語 $L(G)$ と $L(G')$ が $L(G) = L(G')$ であり，かつ $\forall \omega \in L(G), p_\omega(\omega|G) = p_\omega(\omega|G')$ であるとき G と G' は等価であるという．

[定理 2]　任意の曖昧な確率的正則文法 G に対し，これと等価でかつ無曖昧な確率的正則文法 G' は必ずしも存在しない．

[証明]　曖昧な確率的正則文法として図 1 のように状態が増殖するような非決定性オートマトン (NFA) を考える．ノード q_i は状態集合 Q の一つの要素であり，状態を表している．状態遷移図の一つのパスが文法の生成規則の一つに対応しており，確率的文法の場合はパス上に対応する確率が与えられている．状態 q_0 からの遷移について見てみると，生成規則は以下のようになる．便宜上，非終端記号として状態の記号 q_i をそのまま用い，それぞれの生成規則の確率は P_i で表記する(注7)．

$$P_1 : q_0 \to 0q_0$$
$$P_2 : q_0 \to 0q_3$$
$$P_3 : q_0 \to 1q_0$$
$$P_4 : q_0 \to 1q_1$$

1 本のパスが決まればそれを受理する語が一つ決まるが，曖昧な文法の場合では一つの語を受理するためのパスが複数存在する．曖昧な文法を無曖昧な文法で表すために，DFA と NFA の等価性を証明する手法を利用する．NFA の状態の集合 Q に対しそのべき集合を新たな状態の集合 Q' とする．すなわち $Q' = 2^Q$ とする．図 1 の例で NFA と等価な DFA を構成した場合，状態 q_0 から入力 0 が与えられると $\{q_0\} \to \{q_0, q_3\}$ という状態遷移をする．この状態に遷移する新しい確率を $P'(\{q_0\} \to 0\{q_0, q_3\})$，状態 q_0 から入力 1 による遷移の確率は $P'(\{q_0\} \to 1\{q_0, q_1\})$ と表記する．このとき導出の確率として $p_\omega(00)$ と $p_\omega(01)$ はそれぞれ

(注5)：$|\cdot|$ は記号列の長さを表す．

(注6)：この仮定により一般性が失われることはない．

(注7)：$P_1 + P_2 + P_3 + P_4 = 1$ である．

$$p_\omega(00) = P'(\{q_0\} \to 0\{q_0, q_3\})P'(\{q_0, q_3\} \to 0\{q_0, q_3, q_4\}) \tag{6}$$

$$p_\omega(01) = P'(\{q_0\} \to 0\{q_0, q_3\})P'(\{q_0, q_3\} \to 1\{q_0, q_1\}) \tag{7}$$

である．しかし共通のパラメータとなる $P'(\{q_0\} \to 0\{q_0, q_3\})$ は本来 P_1 と P_2 という2つのパラメータ情報を持ち，$\omega = 00$ と $\omega = 01$ の導出において別の値とならなければならない．このため，NFA で決められた語の確率分布で無曖昧な文法では表現できないものが存在する．よって題意が示された． □

確率的正則文法に基づく情報源についての性質が示されたので，次に情報源が既知の場合の符号化アルゴリズムを考える．アルゴリズムは情報源系列を構文解析し，導出木を生成する部分とこの導出木を確率に従って2値符号化する部分の二つからなる．確率からの2値符号化法としては，ハフマン符号や算術符号を用いることして，ここでは情報源系列からその系列の確率を求める部分に焦点をあてる．

定理2から確率的正則文法に基づく情報源の語の生成確率を計算する場合，曖昧な文法について考慮した手続きを考える必要がある．このとき以下の定理を示すことができる．

[定理 3] 確率的正則文法 G において，語 $\omega \in L(G)$ が生成される確率を求めるための計算量は $O(|\omega||\Sigma_N|^2)$ である(注8)．

[証明] 正則文法の左線形性を利用し，語 ω の第1ビットを入力とするような開始記号 S からの生成規則を探索する．正則文法の定義から生成規則の左辺に現れる非終端記号は1文字，右辺に現れる非終端記号は1文字以下であるので，生成規則集合 R の中で終端する規則以外の総数は高々$|\sigma_N|^2$ である．この探索を語 ω の第1ビットから最終ビットまで実行することになるため総計算量は $O(|\omega||\Sigma_N|^2)$ である． □

この定理により，確率的正則文法に基づく情報源の情報源系列 ω の生起確率は線形時間で計算できることが示された．

4. 確率的文脈自由文法

形式文法において正則文法を包含するより一般的な文法として文脈自由文法がある．この文法はプッシュダウンオートマトンと等価であることが示されており，プログラム言語やそのコンパイラの構成などにおいて正則文法では記述できない表現を含んでいるため，実用上重要な文法である．このことからも計算論的情報源として確率的文脈自由文法に基づく情報源モデルを考えることは興味深い．

形式文法の階層構造と同じく，確率的文脈自由文法に基づく情報源は確率的正則文法に基づく情報源を真に包含する．ここでは，情報源既知の符号化として，情報源系列の生起確率の計算量を示し，これが確率的正則文法に基づく情報源の場合の計算より複雑であることを示す．

[定理 4] 確率的文脈自由文法 G において，任意の語 $\omega \in L(G)$ が導出される確率を求めるための計算量は $O(|\omega|^3|\Sigma_N|^3)$ である．

[証明] 文脈自由文法の表記方法として Chomsky 標準形(注9)がある．一方，任意の確率的文脈自由文法でも Chomsky 標準形に変換可能であることが証明されている [1] ため，ここで扱う文法も Chomsky 標準形であると仮定することができる．Chomsky 標準形の文脈自由文法に対し，列 ω の所属性判定アルゴリズムとして CYK アルゴリズム [15] がある．これを語の生起確率の計算手続きに応用することを考える．情報源系列 ω の第 i ビットから長さ j ビットの部分列を ω_i^{i+j-1} と表記する．すなわち，$\omega_i^{i+j-1} = \omega_i\omega_{i+1}\cdots\omega_{i+j-1}$ である．文法 G の生成規則の中で部分列 ω_i^{i+j-1} を導出する非終端記号の集合を v_i^j と表す．ω が与えられたとき R について長さが1の部分列を導出する非終端記号の集合 v_i^1 から再帰的に v_i^j の要素を決定する．CYK アルゴリズムでは v_i^j の要素は生成規則の左辺である非終端記号であるが，$p_\omega(\omega)$ を求めるアルゴリズムではその非終端記号からの導出確率を要素とするように変更する．この確率も再帰的に計算することができる．このときの計算量を考える．すべての v_i^j の個数は高々$|\omega|^2$．また各 v_i^j の要素となりうる生成規則を見つけるための計算量は $O(|\omega|)$ である．さらに v_i^j の中の確率を計算するための乗算の組み合わせ数は $O(|\Sigma_N|^3)$ である．よってすべての計算量は $O(|\omega|^3|\Sigma_N|^3)$ である． □

文脈自由文法は本質的に曖昧であるため，曖昧な文法も考慮しなければならない．次の系は無曖昧な文法に制限した場合でも導出の確率計算が簡単にはならないことを示している．

[系 1] 無曖昧な確率的文脈自由文法 G において，任意の語 $\omega \in L(G)$ が導出される確率を求めるための計算量は $O(|\omega|^3|\Sigma_N|^3)$ である．

定理4は確率的文脈自由文法に基づく情報源が確率的正則文法に基づく情報源を真に包含していることを意味している．またこの二つの情報源の中間に位置するモデルとして無曖昧性の制約条件を加えたものを定義できるが，語の導出確率の計算量では差異は無い．しかし，生成規則の確率パラメータが未知であるような推定問題のときには，曖昧性が大きな影響を与えることになり，情報源モデルを考える場合に曖昧性の制約条件は重要な役割を果たす．その一つの示唆として以下の定理が示されている．

[定理 5] (Soule74 [12]) 文法 G に対し，導出エントロピー H_d と文形式エントロピー H_s において以下の不等式が成立する．

$$H_s(G) \leq H_d(G). \tag{8}$$

ただし，等式は文法 G が無曖昧である場合に成り立つ．

5. 確率的文法に基づく情報源に対するユニバーサル符号

情報源が既知の場合の符号化について述べたが，ここでは情報源のパラメータが未知の場合のユニバーサルな符号について考える．ユニバーサル符号には様々なものが提案されているが，これらのほとんどは文法に基づく符号として考えることが可能

(注8)：$|\cdot|$ は $\cdot$ が集合の場合にはその要素数を表す．

(注9)：生成規則の形が $A \to BC$ または $A \to a$ のいずれかであるような文法である．ただし，$A, B, C \in \Sigma_N, a \in \Sigma_T$ とする．

である[7]．文法に基づく符号は情報源系列 $x \in \Sigma_T^*$ に対して，

$$L(G_x) = \{x\} \tag{9}$$

なる形式文法 G_x を構成し，この文法を算術符号などによって符号化する方法である．復号器はこの文法から言語 $L(G)$ を生成することが可能なため，式 (9) より情報源系列 x を復号できる．ただし文法を符号として利用するためには一意復号性などいくつかの条件を満足する必要がある．このため文法の性質に関する定義をする．

[定義 8] 以下の条件を満足するとき文法 G は許容可能 (admissible) [3] であるという．

(i) G が決定性文法である．

(ii) G の任意の生成規則 $u \to w$ において $w \neq \lambda$ (注10)．

(iii) $L(G) \neq \phi$．

(iv) G は生成規則の中に一度も現れない記号を持たない．

Kieffer らは漸近的にコンパクトな文法に基づく符号のクラスを提案している[3]．以下にこのクラスの定義をする．

[定義 9] G_x を文脈自由文法とし，$|G_x|$ を文法の生成規則の右辺の長さの総和を表すものとする．情報源系列 x から文法 G_x への符号化 $x \to G_x$ において以下の条件を満足するとき，この符号は漸近的コンパクトであるという．

(i) G_x は無曖昧である．

(ii) G_x は許容可能である．

(iii) 非終端記号は順序づけがなされ，導出においてこの順序に従って出現する．

(iv) $\lim_{n\to\infty} \max_{x\in\Sigma_T^n} \frac{|G_x|}{|x|} = 0$

漸近的コンパクト性を持つ符号のクラスを C_{ac} と表記する．このクラスには LZ78 [8] や MPM アルゴリズム [4] などの実用的な符号化法が含まれる．この C_{ac} に含まれる符号は有限状態情報源についてユニバーサル性が明らかにされている．以下に有限状態情報源の定義とユニバーサル性についての定理を示す．

[定義 10] k を正整数とする．以下の式を満足するような要素数が k の状態集合 Q，初期状態 q_0，非負の実数 $\{p(q, x|q')|q, q' \in Q, x \in \Sigma_T\}$ が存在するとき，p_x を k 次の有限状態情報源という．

$$\forall q' \in Q, \quad \sum_{q,x} p(q, x|q') = 1 \tag{10}$$

$$\forall x_1 x_2 \cdots x_n \in \Sigma_T^+,$$

$$p_x(x_1 x_2 \cdots x_n) = \sum_{q_1, q_2, \cdots, q_n \in Q} \prod_{i=1}^{n} p(q_i, x_i|q_{i-1}) \tag{11}$$

[定理 6] (Kieffer [3]) 有限状態情報源に対し，C_{ac} の任意の符号はその冗長度が $O(\nu_n \log(1/\nu_n))$ である．ただし，ν_n は $\max_{x\in\Sigma_T^n} \frac{|G_x|}{|x|} = O(\nu_n)$ となるような 0 に収束するような数列である．

この結果から次の系を導くことができる．

(注10)：λ は空列を表す．

[系 2] 確率的正則文法に基づく情報源に対して，C_{ac} の任意の符号はその冗長度が $O(\nu_n \log(1/\nu_n))$ である．

[証明] 有限状態情報源は遷移の条件に確率が付与された NFA である．この NFA に対し，等価な確率的正則文法を構成することが可能である． □

確率的正則文法に基づく情報源についても C_{ac} に属する符号はその冗長度が漸近的に 0 に収束することが示されたが，より一般的な情報源である確率的文脈自由文法に基づく情報源においても，この漸近収束性が成立することを以下の定理で示す．

[定理 7] 確率的文脈自由文法に基づく情報源に対して，C_{ac} の任意の符号はその冗長度が $O(\nu_n \log(1/\nu_n))$ である．

定理 7 の証明ために次の補題を用いる．

[補題 1] (Kieffer [3]) 漸近的コンパクトな文法に基づく符号によって符号化された文法 G_x に対し，以下の式を満足する情報源系列 x の構文解析 π (注11) が存在する．

$$H(G) \leqq H^*(\pi) + |G| \tag{12}$$

ただし，$H^*(\pi) = \sum_{i=1}^{t} \log \frac{t}{m(u_i|\pi)}$ であり，$m(u_i|\pi)$ は π の中の部分列 u_i の出現頻度を表している．また，$H(G) = H^*(\omega_G)$ であり，ω_G は文法 G の生成規則の右辺を連接したものを非終端記号に従って部分列に分解した構文解析である．

[定理 7 の証明] 確率的文脈自由文法に基づく情報源を $p(x)$ とする．このとき x の任意の任意の構文解析 $(u_1, u_2, \ldots, u_t)$ について次式が成り立つ．

$$p(x) \leqq p(u_1)p(u_2)\cdots p(u_t) \tag{13}$$

なぜなら，確率的文脈自由文法の語 x の確率の計算は x を導出するすべての導出 d (あるいは導出木) の確率の総和であるため，部分列ではそれを導出する可能性のある d の組み合わせはさらに増加するからである．長さ r の終端記号列に対し，次式のような確率分布が存在する．

$$p^*(x) = \frac{\alpha p(x)}{r|\Sigma_N|} \tag{14}$$

ただし，α は $\alpha > 1/2$ となる正定数である．H^* の定義より

$$H^*((u_1, u_2, \ldots, u_t) \leqq \sum_{i=1}^{t} -\log p^*(u_i) \tag{15}$$

が成り立ち，式 (13)(14) (15) より次式が得られる．

$$H(G) \leqq -\log p(x) + t(1 + \log|\Sigma_N|) + |G| + 2\sum_{i=1}^{t} \log|u_i| \tag{16}$$

補題 1 と式 (16) の結果より題意を示すことができる． □

(注11)：π は x をある t 個の部分列に分解したものであり，$\pi = (u_1, u_2, \cdots, u_t)$ でこの u_i を連接したものが x である．

6. むすび

本稿で示した計算論的情報源は確率的形式文法に基づいて情報源系列を生成するモデルである．確率的正則文法と確率的文脈自由文法に基づいた情報源について示した性質から，Chomskyの階層構造がこの計算論的情報源でも保存されることが示せた．これは情報源系列からその生成過程を推定するためのコンプレクシティが，情報源の計算機モデルの複雑さそのものであることを意味する．また，情報源が既知の符号化において二つの情報源における計算量の差異からも階層的な関係を示した．

情報源の確率パラメータが未知であるような場合では，漸近的コンパクトな符号が有限状態情報源について冗長度が0に収束することが示されていたが，本稿ではこれをさらに一般的な情報源である確率的文脈自由文法に基づく情報源においても成立することを示した．

今後，確率的文脈自由文法に基づく情報源における収束の速度を具体的に示す必要がある．また，計算論的情報源として今回示したもの以外の計算機モデル，あるいは形式文法を仮定することができるが，特にLR文法に基づいた情報源モデルについて解析するべきである．一般にプログラム言語やコンパイラなどは決定性文脈自由言語となるように設計されており，LR文法がこの言語を受理する文法であるためである．しかも形式文法の階層において，正則文法と文脈自由文法の中間にLR文法は位置するため，ユニバーサル符号の構成や収束速度を明らかにすることは有用であると考えられる．

さらに計算論的情報源に応じて，文法に基づく符号における文法に適切なモデルを選ぶことにより，従来の符号化法と比較して圧縮率，効率性などについて優れたものを構成できる可能性が生まれるであろう．

謝辞　著者の一人中澤は本研究に際して貴重な御助言をいただきました法政大学の西島利尚先生，早稲田大学の石田崇氏に深く感謝いたします．なお，本研究の一部は科学研究補助金(課題番号12875172)の助成による．

文　献

[1] S. Abney, D. McAllester, F. Pereira, "Relating probabilistic grammars and automata" Proceedings of ACL'99, pp.542-549, 1999.

[2] J.C. Kieffer and E. Yang, "Sequential codes, lossless compression of individual sequences, and Kolmogorov complexity," IEEE Trans. Inform. Theory, vol.42, No.1, pp.29-39, January 1996.

[3] J.C. Kieffer and E. Yang, "Grammar-based codes: A new class of universal lossless source codes," IEEE Trans. Inform. Theory, vol.46, No.3, pp.737-754, May 2000.

[4] J.C. Kieffer, E. Yang, G. Nelson, and P. Cosman, "Universal lossless compression via multilevel pattern matching," IEEE Trans. Inform. Theory, vol.46, No.4, pp.1227-1245, July 2000.

[5] M. Burrows and D.J. Wheeler, "A block-sorting lossless data compression algorithm," SRC Research Report 124, Digital Systems Research Center, May 1994.

[6] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," IEEE Trans. Inform. Theory, vol.23, No.3, pp.337-343, May 1977.

[7] 中澤 真，松嶋 敏泰，平澤茂一，"形式言語と圧縮に関する一考察," 信学技法 IT2001-46, pp.19-24, 2001.

[8] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," IEEE Trans. Inform. Theory, vol.24, No.5, pp.530-536, September 1978.

[9] C.G. Nevil-Manning, I.H. Witten, and D.L. Maulsby, "Compression by induction hierarchical grammars," Proc. IEEE DCC'94, pp.244-253, Snowbird, Utah, USA, March 1994.

[10] C.G. Nevil-Manning and I.H. Witten, "Compression and explanation using hierarchical grammars," The Computer Journal, vol.40, No.2/3, pp.104-116,1997.

[11] C.G. Nevil-Manning and I.H. Witten, "Identifying hierarchical structure in sequences: A linear-time algorithm," J. of Artificial Intelligence Research, vol.7, pp.67-82, 1997.

[12] S. Soule, "Entropies of probabilistic grammars,"" Inform. Contr., Vol.25, pp.57-74, 1974.

[13] M. Li and P.M.B. Vitányi, "An introduction to Kolmogorov complexity and its application," Springer-Verlag, 1993.

[14] M. Li and P.M.B. Vitányi, "A new approach to formal language theory by Kolmogorov complexity," SIAM J. Comput. vol.24, No.2, pp.398-410, 1995.

[15] J.E. Hopcroft and J.D. Ullman, "Introduction to automata theory, languages and computation I," Addison Wesley College, 2000.

[16] J.E.Hopcroft and J.D. Ullman, "Introduction to automata theory, languages and computation II," Addison Wesley College, 2000.

[17] T.A. Welch, "A technique for high-performance data compression," IEEE Computer, vol.17, No.6, pp.8-19, June 1984.