

外2-40

早稲田大学大学院理工学研究科

## 博士論文概要

### 論文題目

Formal Methods in Natural Language Syntax

自然言語構文論における形式的手法

申 請 者

西 哲朗

Tetsuro Nishino

平成 2 年 12 月

生成文法の理論は、N. Chomskyによると、二十世紀に創始され以来、言語学における文法理論の基礎にあると同時に、理論計算機科学の分野においては形式言語理論として発展してきました。近年、人工知能の中心的分野として、自然言語処理（機械翻訳、自然言語インターフェースなど）の重要性が高まり、自然言語に対する計算論的アプローチが活発に研究され始めた。しかし、それらの研究の基礎となる、自然言語構文論の形式的性質（構文解析法、学習可能性、生成能力など）は、あまり解明されていないのが現状である。

本論文では、理論計算機科学における種々の形式的手法を用いて、現代言語学における文法理論の数学的性質を解明すると同時に、それらの理論を計算機上で効率的に実現するための方法論を明らかにする。

ここでは、理論計算機科学における形式的手法とは、形式言語理論、オートマトン理論、計算量理論、帰納的演算論、属性文法理論および計算論的学習理論などを指す。

また、現代言語学における多くの文法理論が提唱されたりするが、本研究では特にR. M. KaplanとJ. Bresnanによると、1982年に提案された語彙機能文法（Lexical-Functional Grammar, 以下LFGと略す）の理論を取り扱う。LFGは文脈自由文法の拡張であり、現在、自然言語処理や認知科学の分野で幅広く用いられる。

一方、文脈自由言語の意味を記述するために、D. E. Knuthによると、提案された属性文法（Attribute Grammar）も文脈自由文法の拡張であり、コンパイラ生成、ソフトウェア工学などの計算機科学の分野で幅広く応用されたりする。

本論文では、ともに文脈自由文法の拡張であるLFGと属性文法の類似性に着目しながら、上記の形式的手法を用いて、それらの文法の数学的性質を解明すると同時に、それらの関係も明確化にする。属性文法を用いてLFGを特徴付けることができるならば、その特徴付けを用いて、LFGに基づく効率の良い構文解析プログラムを計算機上に実現するための指針が得られる。本論文の構成は、概略以下の通りである。

まず、第1章では序論として、言語学、自然言語処理などについて歴史的背景を述べると同時に、理論計算機科学における上述の形式的手法を概観する。さらに、本論文のテーマ、方法論についても詳しく述べる。

第2章では、LFG[3]の理論を概観する。LFGにおけるC構造とF構造という二つの構造を用いて、文法的に正しい文の集合を記述する。C構造は通常の文脈自由型構造とほぼ同じ構造であり、またF構造も、文における文法機能の関係を表現する木構造である。LFGのプロダクションは、文脈自由型プロダクションに、F構造に関する式が付随した形をしてくる。この章では、言語学者が提案したLFG理論を、できるかぎり数学的に厳密に記述することを試みる。

第3章では、論文[3]で提案された形式のLFG理論は、一般的に計算論的に手に

あえない（intractable）ことである。そもそもLFG理論の当初の目標は、効率の良い構文解析と学習が可能であることをある。しかし、MITのR. C. BerwickはLFGの所属問題（与えられた文がそのLFGで生成されるか否かを決定する問題）がNP困難であることを示した[2]。本章では新たに、次の定理を証明する。

定理[8] 与えられたLFGが空集合を生成するか否かを判定する問題（空集合問題）は、決定不能である。

以上の考察から、LFGに何とかの制限を加えない限り、LFG理論の当初の目標は達成できることがわかる。

そこで第4章では、LFGの実際的なサブクラスとしてFrontier-to-Root LFG（以下FRLFGと略す）を提案する。FRLFGと通常のLFGの最も大きな違いは、通常のLFGでは各プロダクションにF構造に関する式が付随しておりたのにに対し、FRLFGにおけるF構造を合成するための代入文が付随しておりとある。この制約によると、C構造内を下から上へ探索しながら、F構造を合成していくことが可能となる（これがFrontier-to-Rootという名前の由来である）。本章ではFRLFGを定義し、その生成のメカニズムについて説明する。さらに、次の定理を証明する。

定理[8]  $L_{CFG} \subseteq L_{FRLFG} \subseteq L_{CSG}$

ここで、 $L_{CFG}$ 、 $L_{FRLFG}$ 、 $L_{CSG}$ はともに文脈自由文法、FRLFGおよび文脈依存文法による生成される言語のクラスである。この定理の前半部分の証明と同様に、2つの言語  $\{a^nb^nc^m | n \geq 1\}$  と  $\{ww^* | w \in \Sigma^*\}$  が FRLFG によると生成できることが示される。よくあるように、これら2つの言語は典型的な非文脈自由言語である。以上のことが、FRLFGの生成能力は、自然言語の構文を記述する上で十分に強力であることがある。

続く第5章では、FRLFG言語の構文解析について考える。まず一般的テーマとして、次の定理が証明される。

定理[8]  $G$  を FRLFG、 $x$  を長さの終端記号列とする。このときの構文解析を  $O(n^3 + d(x, Gr) \cdot n)$  時間で行なうアルゴリズムが存在する。ここに、 $Gr$  は  $G$  の基底文脈自由文法であり、 $d(x, Gr)$  は  $Gr$  における  $x$  の異なる構文木の個数（degree of ambiguity）である。さらに、この定理の特徴をケースとして、次の定理を示される。

定理[8] 基底文脈自由文法が多義的であるFRLFGに対しては、構文解析を  $O(n^2)$  時間で行なうアルゴリズムが存在する。

以上の結果から、FRLFGというサブクラスに限定すれば、計算機上に自然言語に対する効率の良い構文解析プログラムを実現できることがわかる。

第6章では、未知のFRLFGを、学習者が復元を用いて教師から学習する問題について考える。FRLFGはC構造からF構造への変換を記述していることができるが、この変換を実現する木変換機が常に存在することが示される。こ

の結果に基づき、教師は未知の FRLFG を本変換機の形で保持しているものと仮定する。このとき、構造的属性算向、構造的等価性算向、構造的出力算向の3種類の算向を用ひて、未知の FRLFG が異体例から学習する多項式時間アルゴリズムが存在することが証明される。第4章で述べた定理より、本章で示す学習アルゴリズムは、文脈自由言語のクラスを真に言語クラスに対する形式文法の、初めてこの多項式時間学習アルゴリズムであることがわかる。また本章で採用した学習方式は、計算論的学習理論における標準的な学習方式<sup>[1, 10]</sup>の自然な拡張である。

第7章では、LFG のいくつかのサブクラスが、ある種の属性文法<sup>[4, 5, 6, 7]</sup>として特徴付けられることがあります。例えば、次の定理を証明する。

定理[9]  $\mathcal{L}_{\text{PRLFG}} \subseteq \mathcal{L}_{\text{SPAG}}$

ここで、 $\mathcal{L}_{\text{PRLFG}}$ ,  $\mathcal{L}_{\text{SPAG}}$  はそれぞれ Polynomial and Restricted LFG および強多項式属性文法(定義は本文参照)によ、2生成される言語のクラスである。このような特徴付けを用ひると、LFG のサブクラスがもつ生成能力の上界を評価することができる。例えば、上の定理がただちに次の定理が得られる。

定理[9]  $\mathcal{L}_{\text{PRLFG}} \subseteq \text{NP}$

参考文献

- [1] Angluin, D., Learning Regular Sets from Queries and Counterexamples, Information and Computation, Vol. 75 (1987), pp. 87 - 106.
- [2] Berwick, R.C., Computational Complexity and Lexical Functional Grammar, American Journal of Computational Linguistics, 8(3-4) (1982), pp. 47 - 109.
- [3] Kaplan, R.M., and Bresnan, J., Lexical-Functional Grammar: A Formal System for Grammatical Representation, In The Mental Representation of Grammatical Relations, MIT Press (1982).
- [4] Knuth, D.E., Semantics of Context-free Languages, Math. Systems Theory (1968).
- [5] Nishino, T., The Intrinsically Exponential Complexity of the  $k$ -visit Property Problem for Attribute Grammars, In Topology and Computer Science, Kinokuniya (1987).
- [6] Nishino, T., Attribute Grammars and Lexical-Functional Grammars, Proc. of International Computer Science Conference '88 (1988), pp. 426 - 433.
- [7] Nishino, T., Attribute Graph Grammars with Applications to Hierarchical Program Chart Editors, Advances in Software Science and Technology, Vol. 1 (1989), pp. 89 - 104.
- [8] Nishino, T., Mathematical Analysis of Lexical-Functional Grammars, Proc. of Seoul International Conference on Natural Language Processing (1990).
- [9] Nishino, T., Relating Attribute Grammars and Lexical-Functional Grammars, To appear in Information Sciences: An International Journal.
- [10] Sakakibara, Y., Learning Context-Free Grammars from Structural Data in Polynomial Time, Proc. of the First Workshop on COLT, Morgan Kaufmann (1988).