様式 C-19

科学研究費補助金研究成果報告書

平成 21 年　6 月 18 日現在

研究種目：基盤研究（C）
研究期間：2007～2008
課題番号：１９５０００１９
研究課題名（和文）　型つきラムダ計算に基づく構文解析・生成の Datalog への帰着

研究課題名（英文）　Reduction of Parsing and Generation to Datalog Through Typed Lambda Calculus

研究代表者
金沢　誠（KANAZAWA MAKOTO）
国立情報学研究所・情報学プリンシプル研究系・准教授
研究者番号：20261886

研究成果の概要：与えられた文からその文法的構造である導出木を求め、さらに導出木にのっとって文の意味表現を求める問題を構文解析と言い、逆に意味表現から導出木を通して文を求める問題を文生成と言う。文、意味表現、導出木、文法規則などを型つきラムダ計算を使って表現することを通して、構文解析と文生成の両方の問題を、一種の関係データベースに対する問い合わせと見なすことができることがわかった。問い合わせは、Datalog という問い合わせ言語で表現できる。このことから、構文解析と文生成の両方の問題に対して、計算量理論上の位置づけと、統一的な手法による効率的アルゴリズムを得ることができた。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	1,100,000	330,000	1,430,000
２００８年度	700,000	210,000	910,000
年度			
年度			
年度			
総　計	1,800,000	540,000	2,340,000

研究分野：数理言語学
科研費の分科・細目：情報学・情報学基礎
キーワード：形式言語理論、型つきラムダ計算、計算言語学、構文解析、文生成、Datalog

１．研究開始当初の背景

(1) 形式言語理論や計算言語学において、文脈自由文法よりも高い記述力を持つが多項式時間で解析できるような文法形式が多く考案されて来たが、それらのうちの多くが持つ共通の特徴は文脈自由文法で規定できるような導出木の概念を持つことである。2001 年に de Groote (2001)によって多くの文法形式を統一的に捉える枠組みとして、**抽象的範疇文法**(Abstract Categorial Grammar, 以下 ACG)が提案されていた。ACG は、通常の文法の導出や規則にあたるものを**線形ラムダ項**によって表現し、線形ラムダ項の集合を定義する。文字列や木は線形ラムダ項によって自然に表現できるので、ACG は文字列言語を定義する文法や木言語を定義する文法の一般化になっている。導出の形がλ抽象を含まない項（すなわち木）に制限された**２階 ACG** は、文脈自由文法(CFG)、木接合文法(TAG)、線形文脈自由木文法(Linear CFTG)、多重文脈自由文法(MCFG)、多成分木接合文法(MCTAG)などを自然に表現できることが、de Groote (2002), de Groote and Pogodalla (2004), Yamada (2005)の研究で明らかにな

っており、また、de Groote の学生だった Salvati は、2005 年の学位論文で２階 ACG によって定義できる線形ラムダ項の集合が計算量クラス P に属することを証明し、2 階 ACG に対する Earley 流の構文解析アルゴリズムを与えた。このことから、例えば、ACG による TAG の表現を通して TAG に対する Earley 流構文解析アルゴリズムを引き出すことができる。一般の ACG の形式的性質についてはいまだに謎が多く残っているが、この時点で２階 ACG に対する数学的理解はある程度進んでいた。しかし、Salvati の P への所属の証明と構文解析アルゴリズムは必ずしも理解しやすいものだとは言えなかった。

(2) 本研究の研究代表者である金沢は、すべての $n \geq 2$ について n 階 ACG によって定義できる文字列言語のクラスが代入に関して閉じた full AFL をなすことを証明していた(Kanazawa 2006)が、特に、正規言語との共通部分に関する閉包性の証明から、2 階 ACG の言語が P に属するという Salvati の定理の別証明を得ることができた。Salvati は、学位取得後､金沢のもとでポスドクとして ACG の研究を続けていたが、新しい成果として、２階 ACG の定義する文字列言語が決定性木歩行オートマトンの出力言語と一致することを証明し、この結果２階 ACG と多重文脈自由文法が文字列生成能力において等価であることを示した(Salvati 2006)。金沢はこの Salvati の結果にヒントを得て、自身の P への所属の証明を応用することによって、２階 ACG の定義するラムダ項の集合の認識問題が、文法によって定まる Datalog プログラムと入力ラムダ項によって定まるデータベースに対する問い合わせに帰着することができることに気づいた。これはよく知られている CFG の確定節文法による表現の一般化である。Datalog の問い合わせの評価は、プログラムを固定した場合、データベースのサイズに関して多項式時間で計算できることが知られており、これから２階 ACG の定義する集合の P への所属がただちに帰結する。また、Datalog や論理プログラミング一般に対して知られている効率的なアルゴリズムが２階 ACG の定義する集合の認識問題に応用できることも明らかになった。

(3) de Groote (2001)が定義した ACG においては、文法で扱うことのできるラムダ項はλ抽象が常にちょうど１つの変数出現を束縛する線形ラムダ項に制限されていた。この制限のもとで semilinear な言語を定義する CFG, MCFG, Linear CFTG, TAG, MCTAG などの文法フォーマリズムが自然に表現できるが、一方、一般に semilinear でない言語を定義する IO-文脈自由木文法(IO-CFTG)、並列多重文脈自由文法(PMCFG)や､モンタギュー意味論を表現するためには線形でないラムダ項を許すような ACG の拡張（非線形 ACG）が必要とされた。モンタギュー意味論を備えた文法に対する文生成の問題は、文として実現可能(surface realizable)な意味表現の集合を定義する文法に対する構文解析の問題と理解することができるため、非線形 ACG は構文解析と文生成を統一的に捉える枠組みを提供する。

金沢は､2 階 ACG の Datalog による表現が、原子型の変数に限ってλ抽象が２つ以上の変数出現を束縛することを許す**ほとんど線形な**ラムダ項のみを使った**２階疑似線形 ACG** に拡張できることに気づき、学会 LENLS (2006)における招待講演で発表するとともに、Datalog への帰着の正しさの証明を書き上げた。これにより、２階疑似線形 ACG で表現可能な構文解析・文生成の問題が計算量クラス P に属することがわかった。IO-CFTG とモンタギュー意味論のかなりの部分が２階疑似線形 ACG で表現できる。

２．研究の目的

(1) ２階疑似線形 ACG の認識問題が属する計算量クラスをより正確に同定すること、具体的には P の部分クラスである **LOGCFL** に属するかどうかを明らかにする必要があった。文脈自由な導出木を持つ通常の文法の場合は、認識問題が LOGCFL 完全であることが知られている(Engelfriet 1986)。２階疑似線形 ACG についても同様に認識問題が LOGCFL に属することが予想された。このことを証明することが一つの目的であった。

(2) Datalog への帰着から、Datalog や論理プログラミング一般について知られている効率的な評価手法（magic-sets rewriting, OLDT resolution, Earley deduction など）が構文解析と文生成の問題に応用できることが帰結する。これらの手法から具体的にどのようなアルゴリズムが得られるかを確かめ、最適化の方法を検討することがもう一つの目的であった。

３．研究の方法

(1) まず､2 階疑似線形 ACG に対する認識問題が LOGCFL（何らかの文脈自由言語に logspace で帰着できる問題のクラス）に属することを証明する。LOGCFL に属する問題に対しては高速な並列アルゴリズムが存在するから、これは実際的なアルゴリズムの研究にも役立つ。

特別な場合として、２階 ACG の認識問題が LOGCFL に属することを示すことは研究開始前にすでにできていた。２階 ACG の場合、入力ラムダ項に対応するデータベースとして、入力ラムだ項の中の定数記号の出現をそれぞれ別の自由変数で置き換えて得られ

るラムダ項に対する最も一般的な型づけ(principal typing)に対応するものを用いることができ、これは logspace で計算できる。また、すべての２階線形 ACG について文脈自由文法における ε 規則の除去に対応する変形をほどこすことができる(Kanazawa and Yoshinaka 2005)が、この変形の結果から得られる Datalog プログラムと上の方法で得られたデータベースに対する問い合せは、データベースのサイズに関して多項式のサイズの証明木を持ち、このことは答えが yes になる問い合せの集合が LOGCFL に属することを含意する(Ullman and van Gelder 1988, Kanellakis 1988)。

この論法を２階疑似線形 ACG に適用しようとした場合に問題となるのは入力ラムダ項から対応するデータベースを計算する部分であった。文法が線形でないラムダ項を含む場合、入力ラムダ項から対応するデータベースを得る上で、入力ラムダ項に β-簡約するもっともコンパクトなほとんど線形なラムダ項を求める必要があるが、この計算が logspace でできるかどうかが自明ではなかった。有向グラフの congruence closure を求めるアルゴリズムなどを応用してこの問題に対する logspace アルゴリズムを得ることを計画したが、ほとんど線形なラムダ項の性質を使ったより直接的なアルゴリズムも検討した。

(2) マジックセット書き換えに基づくボトムアップ式評価、OLDT resolution, Earley deduction などのメモ化のアイデアを使った Datalog や論理プログラミング一般に対する効率的アルゴリズムを構文解析と生成の問題に応用することを検討する。これらの手法から構文解析と文生成の Earley 流アルゴリズムが自動的に得られることになる。TAG の構文解析や文生成に対する既存の Earley 流アルゴリズムと、このように Datalog への帰着を通して得られる Earley 流アルゴリズムを比較することも計画した。

４．研究成果

(1) ２階疑似線形 ACG に対する認識問題が LOGCFL に属することを証明した。ポイントは、与えられたラムダ項 M から、M に β 簡約するもっともコンパクトなほとんど線形なラムダ項 M' に対するもっとも一般的な型づけ(principal typing)を logspace で計算することができることを示すことであったが、M' を明示的に求めずにその principal typing を logspace で求めることができることをほとんど線形なラムダ項の性質を使って示すことができた。

さらに、与えられた文法から ε 規則にあたるものを除去することなしに、対応する Datalog プログラムが多項式サイズ特性を持つことも証明することができた。（Ullman and van Gelder (1988)では、ε 規則を持たない文脈自由文法を表現する Datalog プログラムが多項式サイズ特性を持つことが示されていたが、ε 規則を持たないと言う条件は不要だったわけである。）文法を標準化することなく認識問題が LOGCFL に属するということを証明できるということは、Gottlob et al. (2002)の結果から、与えられた文法に対する構文解析（導出木を求める問題）が functional LOGCFL に属することを含意するため、単なる別証明以上の大きな意味を持つ。

この他、純粋なラムダ計算の問題として、almost affine なラムダ項が negatively non-duplicated な型づけによって特徴づけられることを証明した。

(2) メモ化のアイデアを使った Datalog の効率的評価手法はいろいろな形で定式化されているが、中でもよく知られているのが **generalized supplementary magic-sets rewriting** にもとづくボトムアップ式評価である。文法を表す Datalog プログラムは構文解析における **deduction system** とみなすことができ、Datalog のボトムアップ式評価はチャート構文解析の制御アルゴリズムで置き換えることができる。このような見方をすると、文脈自由文法を直接的に表現する Datalog プログラムを generalized supplementary magic-sets rewriting によって書き換えてできるプログラムは Earley (1970)のアルゴリズムの deduction system とほぼ同一であり、いろいろな文法に対して generalized supplementary magic-sets rewriting を用いて得られる構文解析アルゴリズムは Earley のアルゴリズムの自然な一般化と見なすことができる。

この手法を２階 ACG による表現をとおして TAG の例に適用してみたところ、結果として得られるアルゴリズムは **correct prefix property** を満たさないことがわかった。TAG については、当初、correct prefix property を満たす Earley 流アルゴリズムは効率を損なうと見られていたが、Nederhof (1999)が $O(n^6)$ 時間の Earley 流アルゴリズムで correct prefix property を満たすものを考案していた。これは correct prefix property を満たさない Earley 流アルゴリズムにいくつかのアドホックな修正を施したものであった。

本研究では、MCFG を直接的に表現する Datalog プログラムから generalized supplementary magic-sets rewriting をとおして correct prefix property を満たす Earley 式のアルゴリズムを得る単純で自然な手法を考案した。単に generalized supplementary magic-sets rewriting を適用

したただけでは correct prefix property が満たされないため、この書き換えの前にもう一つの別の書き換えを適用するのである。これはプログラム中の規則に冗長な subgoal を追加するもので、この結果もとのプログラムと等価なものが得られることは自明である。あとはこのプログラムに対するトップダウン式評価が correct prefix property に対応する性質を満たすことを言えば、その後に generalized supplementary magic-sets rewriting を適用して得られるプログラムを使った構文解析アルゴリズムが correct prefix property を満たすことがすぐに帰結するのである。この方法を TAG を表現する 2-MCFG に対して使えば、TAG に対する prefix-correct な Earley 流構文解析アルゴリズムが自動的に得られる。Nederhof の方法と比較してこの方法はより一般的であり、アドホックな要素がなく、正しさの証明が容易である。TAG に対する時間計算量も同じ $O(n^6)$ であるためアルゴリズムの効率性も犠牲にされていない。

今後は、この手法を発展させて**並列多重文脈自由文法**(PMCFG)や、文字列言語を定義する任意の２階疑似線形 ACG に拡張することを計画している。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

① Makoto Kanazawa. A prefix-correct Earley recognizer for multiple context-free grammars. In *Proceedings of the Ninth International Workshop on Tree Adjoining Grammars and Related Formalisms*, pages 49–56. University of Tübingen. 2008. 査読あり。

② Makoto Kanazawa. Parsing and generation as Datalog queries. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 176–183. Association for Computational Linguistics. 2007. 査読あり。

〔学会発表〕（計２件）

① Makoto Kanazawa. A prefix-correct Earley recognizer for multiple context-free grammars. TAG+9, the Ninth International Workshop on Tree Adjoining Grammars and Related Formalisms. Tübingen, Germany. June 7, 2008.

② Makoto Kanazawa. Parsing and generation as Datalog queries. The 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic. June 25, 2007.

６．研究組織

(1)研究代表者

金沢　誠（KANAZAWA MAKOTO）

国立情報学研究所・情報学プリンシプル研究系・准教授

研究者番号：20261886

(2)研究分担者

なし

(3)連携研究者

なし