様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 ２０ 日現在

機関番号：６２６１５
研究種目：基盤研究(C)（一般）
研究期間：2013～2015
課題番号：２５３３００２０

研究課題名（和文）文脈自由木文法の生成する木言語および文字列言語の性質の研究

研究課題名（英文）A Study of Tree and String Languages Generated by Context-Free Tree Grammars

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

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：単純文脈自由木文法(CFTsp)に対して，Chomsky-Schuetzenbergerの定理の自然な拡張となる定理を証明した。Chomsky-Schuetzenbergerの定理がDyck言語を使うのに対して，この定理は，「Dyck木言語」の概念を用いる。応用として，CFTspと等価な樹状指標文法の概念を考案した。さらに，CFTspが生成する文字列言語の特徴づけを得ることもできた。
また，CFTspが生成する文字列言語に対してはOgdenの補題が成り立たないことを証明し，多重文脈自由文法に対してOgdenの補題が成り立つための十分条件を提示した。

研究成果の概要（英文）：I obtained a Chomsky-Schuetzenberger-style representation theorem for simple context-free tree grammars. This theorem uses a notion of Dyck tree language, in contrast to Dyck languages used by Chomsky and Schuetzenberger. As an application of this theorem, I conceived a new grammar formalism called "arboreal indexed grammars", which exactly correspond to simple context-free tree grammars. I also obtained a representation theorem for the string languages of simple context-free tree grammars, which is a natural generalization of Weir's representation theorem for string languages of tree-adjoining grammars.
Further, I showed that an Ogden-style iteration theorem does not hold for the string languages of simple context-free tree grammars, and gave a sufficient condition for a multiple context-free grammar to satisfy an Ogden-style iteration theorem. Every language in Weir's control language hierarchy can be generated by a multiple context-free grammar satisfying this condition.

研究分野：数理言語学

キーワード：単純文脈自由木文法　樹状指標文法　多重文脈自由文法　多次元木　表現定理　Ogdenの補題　Dyck木言語

1．研究開始当初の背景

（1）文脈自由木文法は，木の集合（木言語）を生成する文法形式の一つで，木や木パターンを組み合わせる規則の体系である。木の葉節点のラベルを並べてできる文字列（木の「産出」）を取ることによって，文字列集合を生成する文法と見なすこともできる。

単純文脈自由木文法は，文脈自由木文法の規則に対して，構造の消去や複製を許さないという制限を課したものである。一般の文脈自由木文法が必ずしも抽象的な意味での「文脈自由な文法」の概念に合致しないものであるのに対して，単純文脈自由木文法は，通常の文字列言語に対する文脈自由文法を木言語に自然に拡張したものであると言える。単純文脈自由木文法は，近年，データ圧縮の効率的手法として研究されて来ていた他，理論計算言語学の分野でも注目を集めていた。

（2）自然言語の記述にもっとも広く用いられている文法形式の一つである木接合文法(TAG)は，ランク 1 の単純文脈自由木文法と等価であることが知られていた。自然言語を記述する文法の望ましい性質の一つとして，文法の「語彙化」可能性がある。語彙化とは，すべての文法規則に終端記号が出現するような形に文法を変形することである。長い間 TAG は TAG によって語彙化できることが信じられていたが，Kuhlmann and Satta (2012)はこれが一般には不可能であることを示した。Malletti and Engelfriet (2012)は，TAG はランク 2 の単純文脈自由木文法によって語彙化でき，一般にランク r の単純文脈自由木文法はランク $r + 1$ の単純文脈自由木文法によって語彙化できることを示した。したがって，単純文脈自由木文法は，語彙化可能な文法形式である。

（3）単純文脈自由木文法は，文字列生成能力の観点からは，文字列の組を操作する規則の体系である多重文脈自由文法に対して「入れ子条件」を課したものと等価であることが知られていた(Kanazawa 2009)。長いあいだ，多重文脈自由文法が Joshi (1985)の「弱文脈依存」文法の概念を形式化したものであると考えられていたが， Joshi et al. (1991)によって弱文脈依存文法の記述力を超えるとされた形式言語

MIX = { $w \in \{a, b, c\}^* \mid w$ は a と b と c を同数含む }

が多重文脈自由文法で記述できることが 2011 年に発見される（この結果は後に Salvati 2015 として出版）など，最近になって，弱文脈依存の概念をとらえるには入れ子条件のような制限を課すことが必要であるという考えが生まれて来ていた。

（4）Yoshinaka et al. (2010)が多重文脈自由文法に対して Chomsky–Schützenberger の定理に似た表現定理を証明したが，ここで使われていた「多重 Dyck 言語」の概念は必ずしも通常の Dyck 言語の自然な一般化とは言えないものであった。

（5）研究代表者は，多重文脈自由文法のクラス全体及びその部分クラスに関して，次のような研究を行って来た。

① Earley 型構文解析アルゴリズム　多重文脈自由文法全般に対して，入力文字列中の誤りを即座に検出するという性質(correct prefix property)を満たした Earley 型構文解析アルゴリズムを自動的に生成する手法を考案した(Kanazawa 2008)。

② ポンプの補題　入れ子条件を満たした多重文脈自由文法によって生成される文字列言語のクラス（つまり，単純文脈自由木文法によって生成される文字列言語のクラス，これを yCFTsp と書く）に対して自然な形のポンプの補題が成り立つことを証明した(Kanazawa 2009)。

③ 複製定理　yCFTsp に属する $\{ w\#w \mid w \in L \}$ の形の言語を特徴づけた(Kanazawa and Salvati 2010)。これによって，多重文脈自由文法によって生成される文字列言語のクラスと単純文脈自由木文法によって生成される文字列言語のクラスとの間に大きな乖離があることが明らかになった。

④ MIX 問題の解決　MIX 言語が TAG によって生成できないという Joshi (1985)の予想に肯定的解答を与えた(Kanazawa and Salvati 2012)。

2．研究の目的

多重文脈自由文法と比較して，単純文脈自由木文法は通常の文脈自由文法の性質をより多く保存しているように見える。本研究は，この印象を裏付けるような新たな具体的結果を集めることを目的とした。特に，研究代表者のこれまでの研究の中から浮かび上がって来た次のような問題を解決することを目指した。

（1）単純文脈自由木文法に対する表現定理

Chomsky–Schützenberger の定理は，通常の文字列に対する文脈自由文法の導出木の文字列表現と Dyck 言語を正規言語と準同型写像を使って結びつけるものである。単純文脈自由木文法の導出木を 3 次元木で表すことにより，単純文脈自由木文法の木言語及び文字列言語に対する表現定理を証明する。

（2）yCFTsp に対する時間計算量的に最適な Earley 型アルゴリズム

入れ子条件を満たした多重文脈自由文法に対しては，文法の次数 m に応じて，$O(n^{2m+2})$ の時間計算量を持つ認識アルゴリズムがあることが知られていた(Gómez-Rodríguez, Kuhlmann, and Satta 2010)。しかし，Kanazawa (2008)の手法を入れ子条件を満たした多重文脈自由文法に適用しても，この

時間計算量を持つアルゴリズムを得ることができない。Kanazawa (2008)は TAG に対して時間計算量的に最適な Earley 型アルゴリズムを得る手法を示している。これを一般化して，単純文脈自由木文法に対して $O(n^{2m+2})$の時間計算量を持つ Earley 型アルゴリズムが得られることを示す。

（3）yCFTsp に対するポンプの補題の精緻化・Ogden の補題

yCFTsp に対して，反復される文字列の位置や長さに制限を加えた形のポンプの補題や，Ogden の補題を証明する。

（4）MIX が yCFTsp に属さないことの証明

Salvati (2015)と Kanazawa and Salvati (2012)によって, MIX が次数 2 の多重文脈自由文法によって生成できるが，入れ子条件を満たした次数 2 の多重文脈自由文法によっては生成できないことがわかった。しかし，次数 3 以上の入れ子条件を満たした多重文脈自由文法で MIX が生成できるか，すなわち MIX が yCFTsp に属するかどうかという問題は解決されていなかった。この問題に対する否定的な解答を与えることを目指す。

（5）その他の関連する問題

一般の IO 文脈自由木文法の生成する文字列言語のクラス（これは yCFTsp を真に包含する）と並列多重文脈自由文法の生成する文字列言語のクラスの関係など，周辺の未解決問題に取り組む。

３．研究の方法

フランスの INRIA Bordeaux の Sylvain Salvati を研究協力者として研究を進めた。特に，研究目的の項目（4）と（5）については，お互いが相手を訪問する機会を利用して，共同研究として行った。

まず，平成 25 年度に，方針のはっきり定まっていた研究目的の（1）の項目に重点的に取り組むこととした。同時に, 研究目的（2）と（4）についても Salvati と共同で研究を進めることとし，さらに平成 26 年度以降に研究目的の（3）と（5）に取り組むことを計画した。

研究目的の（2）については予想が誤りであったことが判明したため，途中で方針を変更して別の手法を試みることを計画したが，着手することはできなかった。

４．研究成果

研究目的の（1）については想定通りの成果をあげることができ，またさらにその帰結として単純文脈自由木文法とぴったり対応する指標文法に対する制限を見つけることができた。研究目的の（3）については，予想とは逆の否定的な定理を証明することができ，さらに多重文脈自由文法について当初想定していなかった定理を証明することができた。研究目的の（2）と（4）については意義のある成果をあげることができなかった。研究目的の（5）については，1 つの問題を解くことができたが，論文執筆までには至らなかった。

以下，得られた成果について述べる。番号は研究目的の項目とは一致しない。

（1）単純文脈自由木文法に対する表現定理

通常の文字列言語を生成する文脈自由文法に対する Chomsky–Schützenberger の定理は, 文脈自由言語を Dyck 言語と正規言語，及び準同型写像という 3 つの要素を使って特徴付けるものと一般に理解されているが，もともとの Chomsky の意図は，文脈自由文法の導出木を Dyck 言語を使って表現することであった。この視点を保ちながら，Chomsky–Schützenberger の定理を単純文脈自由木文法に拡張することに成功した。

Dyck 言語の要素は，通常の木を括弧を使って文字列で表現したものと捉えることができる。Baldwin and Strawn (1991)の多次元木の概念を使うと，単純文脈自由木文法の導出木は 3 次元木と見なすことができる。3 次元木を一般化された括弧の概念を用いて 2 次元木で表現すると，「Dyck 木言語」の概念が得られる。本研究では，Dyck 木言語を使って Chomsky–Schützenberger の定理を自然に一般化することによって，単純文脈自由木文法に対する表現定理を得た。この定理は, 単純文脈自由木文法が生成する木言語を特徴付けると同時に単純文脈自由木文法の導出木の集合（3 次元木の集合）を通常の 2 次元木の集合で表現したものでもあるという点で, もともとの Chomsky–Schützenberger の定理の精神を引き継ぐものである。

Dyck 木言語を使って表現された木集合をさらに文字列集合で表現することを考え，単純文脈自由木文法の生成する文字列集合に対しても表現定理を証明することができた。これは，Dyck 言語で用いられる開く括弧と閉じる括弧の対応関係を変更することによって得られる 2 つの Dyck 言語の共通部分を用いて yCFTsp の要素を表現するもので，TAG の生成する文字列言語に対して Weir (1988)が証明した表現定理の自然な一般化になっているものである。

（2）単純文脈自由木文法と等価な線形指標文法の一般化

指標文法は，文脈自由文法の非終端記号の各出現に指標の列を格納したスタックを貼り付け，書き換え規則によってスタックを操作する文法形式である。指標文法は，一般の文脈自由木文法を OI 導出で解釈したものと等価であることが昔から知られている。また，指標文法の変種である線形指標文法は，TAG と等価であることがわかっていた。線形指標文法のスタックを操作する規則は，スタック

に指標をプッシュする規則とスタックから指標をポップする規則とが，開く括弧と閉じる括弧のように対をなして働き，このことが線形指標文法と TAG との等価性の証明で使われている。このことに着目して，指標をプッシュする規則と指標をポップする規則の組み合わせが Dyck 木言語の一般化された括弧のように振る舞う指標文法の変種を考案した。

新しい指標文法の変種を定義する準備として，一般の指標文法に対して通常とは異なる解釈を与えた。一般の指標文法の通常の解釈では，導出木の中でスタックが（最上部を除いて）非終端記号のラベルを持つ全ての子節点に引き継がれるが，線形指標文法では，スタックが引き継がれるのは規則が指定した 1 つの子節点のみである。従って，線形指標文法は指標文法の特殊なケースではなく，類似した別の文法形式であることになる。しかし，指標文法のスタックの伝搬をトップダウンではなくボトムアップに解釈することも可能である。つまり，子節点のスタックは全て同じでなくてもよく，2 つの子節点のスタックは一方が他方の prefix になっていれば良いとし，親節点は子節点のスタックのうち一番長いものを（変化する最上部を除いて）引き継ぐものとするのである。このような解釈を与えると指標文法の導出木は変化するが，変化するのはスタックの部分のみであり，生成される文字列言語にも影響はない。

このように指標文法の規則をボトムアップに解釈すると，線形指標文法は，指標文法のうち，導出木の中で 1 つの指標をプッシュする規則と対応する指標をポップする規則が高々1 つであるという条件を満たしたものと見なすことができる。これを一般化し，導出木の中で 1 つのプッシュ規則と対応するポップ規則の数が m 以下と言う条件を考えると，ランク m の単純文脈自由木文法のクラスと等価な指標文法のクラスが得られる。

上で述べた指標文法に対する制限を導出木に対する大域的な制限ではなく，一つ一つの規則の形に対する制限で表現することもできる。このためには，スタック部分に指標の列ではなく，指標からなる一種の木を格納することが必要になる。この新しいタイプの指標文法を「樹状指標文法」と名付けた。

（3）入れ子条件を満たした多重文脈自由文法に対する Ogden の定理の不成立と「適切な」多重文脈自由文法に対する Ogden の定理

入れ子条件を満たした多重文脈自由文法に対しては，Kanazawa (2009)が自然な形のポンプの補題を証明したが，これを反復される文字列の長さや位置に関する制約を含めた形に強めることができることが予想された。また，研究期間中に，等価な文法クラスに対して Ogden の補題が成り立つことを主張する論文が発表された(Sorokin 2014)。しかし，この論文を精査した結果，証明に誤りがあることがすぐにわかり，その後，具体的な反例を使って Ogden の補題が入れ子条件を満たした多重文脈自由文法に対して成り立たないことを示すことができた。同じ反例は，Kanazawa (2009)でポンプの補題が成り立つことが示されていたもう一つのクラスである次数が 2 の多重文脈自由文法に対しても Ogden の補題が成り立たないことを示すものである。

入れ子条件はポンプの補題の成立を保証するが Ogden の補題を含意しないことが明らかになったが，この結果を得る過程で，Ogden の補題を含意する自然な条件を見つけることができた。この条件を満たした多重文脈自由文法を「適切な」文法と呼ぶ。適切さの条件と入れ子条件はお互いに相手を含意しない関係にある。

多重文脈自由言語の部分クラスに相当することが Kanazawa and Salvati (2007)によって示されていた Weir の制御言語の階層(Weir 1992)に対して，早くから Ogden の補題の一般化が成り立つことが知られていたが(Palis and Shende 1995)，制御言語はすべて適切な多重文脈自由文法で生成できることも示すことができた。したがって，制御言語に対しては Palis and Shende の Ogden の補題と本研究で得られた Ogden の補題の両方が成立することになるが，両者は同じ主張ではなく，反復される文字列やその近傍に対して課す制限が異なる。本研究で得られた Ogden の補題は反復される文字列に含まれるマークされた文字の出現数の合計がある定数を超えないという条件を含んでおり，Palis and Shende の Ogden の補題と比較して，もともとの Ogden の補題のより自然な一般化であると言える。

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　3 件）

① Makoto Kanazawa. Ogden's lemma, multiple context-free grammars, and the control language hierarchy. In Adrian-Horia Dediu, Jan Janoušek, Carlos Martín-Vide, and Bianca Truthe, editors, *Language and Automata Theory and Applications, 10th International Conference LATA 2016*, pages 371–383. Lecture Notes in Computer Science 9618. Springer. 2016. 査読有
DOI: 10.1007/978-3-319-30000-9_29

② Makoto Kanazawa. A generalization of linear indexed grammars equivalent to simple context-free tree grammars. In Glyn Morrill, Reinhard Muskens, Rainer Osswald, and Frank Richter,

editors, Formal Grammar, FG 2014, pages 86–103. Lecture Notes in Computer Science 8612. Berlin: Springer. 2014. **査読有**
DOI: 10.1007/978-3-662-44121-3_6

③ Makoto Kanazawa. Multidimensional trees and a Chomsky–Schützenberger–Weir representation theorem for simple context-free tree grammars. *Journal of Logic and Computation*. Published online June 30, 2014. **査読有**
DOI: 10.1093/logcom/exu043

〔学会発表〕（計　2 件）

① Makoto Kanazawa. Ogden's lemma, multiple context-free grammars, and the control language hierarchy, 10th International Conference on Language and Automata Theory and Applications, Prague, Czech Republic, March 16, 2016.

② Makoto Kanazawa. A Generalization of Linear Indexed Grammars Equivalent to Simple Context-Free Tree Grammars, FG-2014: The 19th Conference on Formal Grammar, University of Tübingen, Tübingen, Germany, August 16, 2014.

〔その他〕
ホームページ等
http://research.nii.ac.jp/~kanazawa/

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

(4)研究協力者
SALVATI, Sylvain
INRIA Bordeaux, France