論文の内容の要旨

論文題目: Collaborative and Corpus-Driven Approaches towards Lexicalized Grammar-based Natural Language Processing
（和訳: 語彙化文法による自然言語処理の実現に向けて ―
共同的かつコーパスに基づくアプローチ）

氏　　名　　吉 永　直 樹

情報抽出や QA システム、機械翻訳といった、知的な言語処理アプリケーションでは、品詞列や係り受けといった Shallow Syntax だけではなく、動詞を述語とした述語項構造のような Deep Syntax や、さらには語彙意味論や構成的意味論などで表現される意味表現のようなより深い文解釈が必要となる。このような Deep Syntax や意味表現を文に対して与えることが出来る枠組の中でも、語彙化木接合文法（LTAG）や主辞駆動句構造文法(HPSG）に代表される語彙化文法は、語彙を中心として言語の統語的側面だけでなく意味的側面も統合的に扱おうとする枠組であり、上記のような実アプリケーションのためのコアのコンポーネントとして有望視されている。

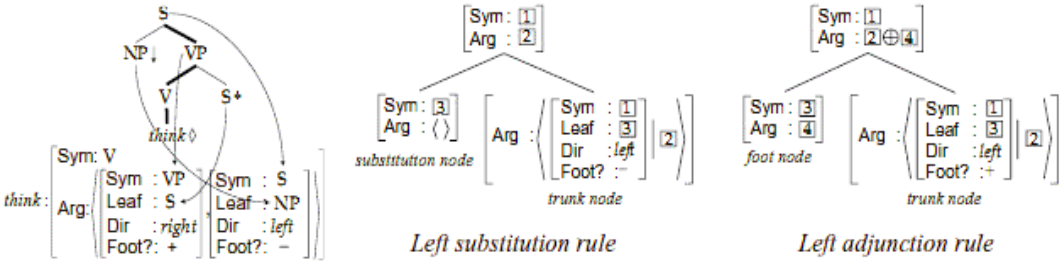
本論文では、この語彙化文法を用いた自然言語処理の実現に向けて、二つのアプローチを提案する。具体的には、文法規則などの静的な文法資源や、語彙化文法一般に適用可能な汎的な言語処理技術（例えば構文解析器、曖昧性解消モジュールなど）を語彙化文法の個々の枠組を越えて共同的に開発を進めるための方法論を提案する。一方で動的にタスクの分野に応じて追補的に必要となる辞書資源については、既存の文法に含まれる言語学的な一般化を考慮することで、対象分野からより正確な辞書資源を獲得する手法を提案する。

語彙化文法は、単語に対する構文的制約と意味的制約を記述した語彙項目を、少量の文法規則で組み合わせることで、文に対し構文構造・意味表現を与える文法枠組である。構文的制約としては、例えば単語の共起に関する構文知識（動詞の下位範疇化要素（目的語・補語など）の品詞情報）などが、意味的制約としては単語の共起に関する意味知識（例えば drink が目的語に液体を取るといった選択制約）などがある。このように定義される語彙化文法を実際のアプリケーションの文脈に応用しようとすると、辞書項目のデータ構造・文法記述が複雑・詳細であるため、2) 高効率な構文解析器を実現する必要がある上に、2) 広範かつ詳細な辞書資源を得ることが困難という問題がある。

本論文の前半では、語彙化文法の枠組の間にまず文法変換アルゴリズムを開発し、その文法変換を個々の枠組の表層的な違いを捨象する手段として用い、語彙化文法に対する汎的な言語処理技術を開発する方法論について述べている。これまで、文法変換は、個々の枠組みの間の辞書リソースを共有することを目的として提案されてきたが、本研究ではこ

れに加えてさらに、文法変換によりまず、ある文法枠組の文法を他の文法枠組みの強い意味に等価な文法に変換することで、1) 等価な文法を仲立ちとして文法以外の構文解析器、曖昧性解消モジュール、文法開発環境などの言語処理技術が共有できることを示す。さらに我々は、2) それらの技術を等価な文法を用いて比較することで、枠組非依存の一般的な技術（例えば構文解析技術）に対する深い洞察を得ることができ、その洞察に基づき既存の言語処理技術を改善できることを示す。ここで言う等価な文法というのは、二つの文法が、同じ文に対し一対一に変換可能な構文解析結果を返すことを指す。我々は LTAG から HPSG スタイルの文法への文法変換を提案・実装し、文法リソースの共有の観点と、言語処理技術の比較・検討の観点からそれぞれ実験を行った。

以下でまず、LTAG 文法から HPSG スタイルの文法への文法変換について述べる。我々の提案した文法変換は、1) LTAG の語彙項目(木構造, 図吉永直樹画像.png 左上)を HPSG の語彙項目（素性構造, 図吉永直樹画像.png 左下）に変換し、2)LTAG の文法規則を模倣する HPSG の文法規則を定義する、という 2 項目からなる。1 について、LTAG と HPSG とでは文法的制約の局所化、すなわち複数の単語にまたがる文法的制約を、どの単語の語彙項目の文法的制約として記述するかという点で違いがあり、文法の語彙項目同士が一対一に対応しないため単純な変換ができない。そこで我々は、HPSG の語彙項目に一対一に対応する木構造(canonical tree)を定義し、LTAG の語彙項目に記述された文法的制約を、木構造を canonical tree に変換することで、HPSG の観点から見た文法的制約として捉え直すという方針を採った。こうして得られた canonical tree は、一つの単語に対する HPSG 的観点から見た構文的・意味的制約を含むため、葉ノードの品詞ラベルを語の下位範疇化要素と捕らえスタックに保存することで HPSG の語彙項目に変換できる（図 1 左)。さらに文法規則をこのスタックに保存された構文的・意味的制約を操作するように定義することで、LTAG の文法規則を模倣する（図 1 中央・右)。実験として、米ペンシルバニア大学で開発されている大規模 LTAG 文法を変換し、等価な HPSG スタイルの文法が得られることを確認した。



我々は次に、語彙化文法という枠で一般的に有効な言語処理技術を構築することを目的として、既存の LTAG 文法と文法変換で得られる等価な HPSG スタイルの文法を利用し、LTAG と HPSG という異なる文法枠組で開発された構文解析器の構文解析速度の違いを比

較・分析した。このような比較実験は、我々が初めて提案した等価性を保証する文法変換により実現可能となった。実験では、上記の実験で得られた等価な LTAG および HPSG スタイルの文法を用い、動的計画法と CFG フィルタリングと呼ばれる構文解析手法について LTAG と HPSG とで別々に設計された構文解析器を比較した。その結果、HPSG の構文解析器がチャート法（13.5 倍）、CFG フィルタリング（30 倍～230 倍）共に高速であることを示し、さらに、その実装方法の差異を分析することで、LTAG の構文解析器の改良法についても考察した。これにより、構文解析手法の開発について、文法枠組の違いを越えて共同的にアプローチすることができるようになったと言える。

本論文の後半では、既存の文法リソースと統合するのに十分な正確さ備えた語彙化文法の文法リソースを、生コーパスから獲得する手法を提案する。我々は文法変換を用いて語彙化文法を実アプリケーションに用いる際の必要となる言語処理技術を共同的に開発する方法論を提示したが、実際に語彙化文法を実アプリケーションに用いる際には、既存の人手で書かれた文法は文法の広範性について問題が残る。近年これに対し、語彙化文法の語彙項目を括弧つきコーパスから自動獲得するという流れと、既存の文法をコーパスから獲得した文法的知識（動詞の下位範疇化フレーム）で増強する流れで研究が進められている。しかしながら、文法の広範性と詳細性はトレードオフの関係にあり、前者の立場で研究を進めると、文法の詳細性、あるいは言語学的な妥当性という点で問題が生じ、後者の立場で研究を進めると、得られる文法的知識と既存の文法の一貫性の点で問題が生じる。また、人手で記述された文法知識、また、有限の注釈つきコーパスから得られた文法知識は、正確ではあるものの、一般的に広範でない事が指摘されている。従って、注釈無しの大量に利用できるコーパスから獲得した文法リソースが、実際の対象ドメインを考えたときの言語処理では必要不可欠である。しかしながら、近年行われた研究では、生コーパスから獲得された確度の低い辞書項目により文法を増強した場合、構文解析器の性能を著しく低下することが報告されている。

我々は、我々はより信頼性の高い知識を生コーパスから獲得することを目的として、対象文書から既存手法により獲得した構文的知識（動詞の下位範疇化フレーム）を、既存の文法の辞書に含まれる言語学的な一般性の元で検証することで、コーパスから獲得された文法的知識の質を既存の文法と一貫性を保つように改善する手法を提案した。提案手法では、既存の語彙化文法の辞書に含まれる動詞と対象文書から下位範疇化フレームを獲得した動詞に対し、それらの下位範疇化フレームの共起をクラスタリングすることで、誤って獲得された構文的知識を除去し、信頼性の高い構文的知識を獲得することを可能とした。実験として、携帯電話の会話文書から獲得された動詞の構文的知識を、既存の LTAG 文法および HPSG 文法の辞書を用いて改善することに成功した。