論 文 の 内 容 の 要 旨

論文題目　From Linguistic Theory to Syntactic Analysis:
Corpus-Oriented Grammar Development and Feature Forest Model
（言語理論から統語解析へ：コーパス指向の文法開発と素性森モデル）

氏　名　宮尾　祐介

本論文では，実世界テキストの自動統語解析を行うシステムを構築することを目指す．ここで統語解析とは，主辞駆動句構造文法(HPSG)などの統語理論により言語学的に理論付けされた統語構造を計算すること指す．自然言語の意味を計算するためには統語構造は不可欠であり，実際，様々な知的自然言語処理，例えば質問応答，対話システム，テキストマイニングなどにおいて需要が高まっている．しかしながら，新聞などの実世界テキストを解析できる性能の統語解析器は現在のところほとんど存在しない．

本課題には，二つの大きな障壁が存在する．一つは文法のスケーラビリティの問題である．統語解析の理論的研究は現在までにさかんに行われてきたが，実世界のテキストを解析できるまでの大規模な文法を開発するのはほぼ不可能であった．これは文法を拡大していく際の本質的な問題に起因すると考えられる．それは，文法を拡大するにつれて，文法の一貫性の維持が困難になるという問題である．現代の統語理論は語彙化文法と呼ばれる枠組みが主流で，文の統語構造をプリンシプルと語彙項目の様々な組み合わせで説明する．プリンシプルは統語構造が一般的に満たすべき制約を記述したもので，語彙項目は単語特有の性質（品詞や下位範疇化フレームなど）を記述したものである．しかし，文法開発者がプリンシプル・語彙項目のすべての組み合わせ可能性を考慮してこれらを実装するのは非常に困難である．特に，実世界のテキストを解析するためには大量の単語に対して語彙項目を適切に記述する必要があるが，各単語の可能な全ての使われ方を網羅し，かつプリンシプルや他の単語の語彙項目との一貫性を保ちつつ文法を拡大していくのはほぼ不可能である．それに加えて，統語理論は詳細な統語情報（時制や格など）の複雑な関係を扱うため，一貫性の維持はなおさら困難になる．

この問題に対して，本論文では，文法開発の新しい方法論を提案する．本手法，**コーパス指向の文法開発**では，語彙項目を開発する代わりにツリーバンクを作成する．ツリーバンクとは，実世界の文に対して，その文の解析結果を人手で与えたデータである．すなわち，統語解析の正解例のデータであり，統語構造の実例のデータである．まず，文法開発者はターゲットの統語理論（本論文では HPSG）に則ったプリンシプルを定義する．次に，既存の言語リソースを利用し，ターゲットの統語理論に則ったツリーバンク（HPSG ツリーバンク）を作成する．例えば，文解析の研究で広く利用されている Penn Treebank は文脈自由文法レベルの情報しか与えないが，これを変換することにより，HPSG 理論に基づくツリーバンクを得ることができる．文法開発者の主な仕事は，この変換過程をコントロールすることとなる．その際，プリンシプルをツリーバンクに適用することによって，ツリーバンク中のエラーや非一貫性はプリンシプルの違反という形で自動的に検出することができる．つまり，文法開発者は非一貫性の原因を簡単に同定することができ，それを修正していくことでより高品質なツリーバンクを作り上げていくことができる．すなわち，本方法論における文法開発とは，統語理論に則ったプリンシプルを満足するようにツリーバンクを作成していくことを意味する．これは，文法の一貫性を実テキストの解析例の範囲

内で保証するということになる．十分な大きさのツリーバンクを開発することができれば，語彙項目はツリーバンク中の統語解析木の終端ノードを集めることで自動的に得られる．本手法により，文法の一貫性の制御が容易になり，また既存の言語リソースが再利用できるため，大規模な文法を効率的に開発できると期待される．

実世界テキストの統語解析におけるもう一つの障壁は，自然言語の選好性のモデル化である．統語理論の研究は主に構造的規則性に焦点が当たっていたため，選好性のモデル化はあまり議論されていなかった．しかし，選好性は自動統語解析には必要不可欠である．なぜなら，実アプリケーションは曖昧性が解消された，もしくはランク付けされた解析を要求するからである．文脈自由文法に基づく文解析では統計・確率モデルが一定の成功を収めたため，本研究でも確率モデルの構築を目指す．

しかしながら，HPSG など語彙化文法は，型付き素性構造などの複雑なデータ構造で記述されているため，既存の確率モデルは適用できない．既存の自然言語処理技術では，文全体の構造を部分構造に分解し，その部分構造間の統計的独立性を仮定し，部分構造の確率モデルを推定する，というアプローチをとる．例えば，文の品詞列は単語ごとの品詞，文全体の構文木は一段の分岐に分解され，分解された部分構造の確率を推定し，全体構造の確率は部分構造の確率の積として定義する．これらの手法は，対象問題の構造が木構造や格子構造であることを前提としている．型付き素性構造は任意のグラフ構造であるため，この手法は適用できない．

この問題に対する解決策として，本論文では **feature forest モデル**を提案する． Feature forest とは, 指数関数的数の木構造をパックした構造で表現するデータ構造である. Feature forest モデルは, feature forest の上に定義された最大エントロピーモデルである．本論文では，feature forest 中の木の確率モデルのパラメータを，feature forest を展開することなく推定する動的計画アルゴリズムを提案する．これにより，どのような構造の確率事象であっても，feature forest で表現されていれば統計的独立性の仮定なしに確率モデルを推定することができる．また，本論文ではさらに HPSG の統語構造や述語項構造が feature forest で表現できることを示す．これらにより，HPSG に基づく統語解析のための確率モデルを構築する方法が完成する．

本研究で提案した手法を実装し，HPSG に基づく英語統語解析器を開発した．英語の新聞記事のツリーバンクである Penn Treebank をテストデータとして，統語解析の実験を行い，本統語解析器の実用性および提案手法の有効性を示す．まず，Penn Treebank を変換することで大規模な HPSG ツリーバンクを作成した．そこから語彙項目を獲得し，英語の大規模 HPSG 文法を開発した．文法の実テキストに対する被覆率を測定した結果，99.7%の文に対して何らかの解析結果を出力することができ，またそのうち 84.1%の文に対しては正解の統語構造を含む解候補を出力することを確認した．この被覆率は現在までの統語解析器では達成し得ない高い性能を示している．また，feature forest モデルを適用し，HPSG に基づく統語解析の曖昧性解消モデルを開発した．曖昧性解消の結果，述語項関係の精度を測定したところ，87.69%/87.16%（適合率/再現率）の精度を達成した．その他にも，実テキストの解析における HPSG 統語解析器の性質を示すため，様々な実験結果を提供する．

本論文では，言語理論に基づく統語解析を実用レベルのシステムとするための基本的な方法論を提供し，その有効性を実証することができた．実テキストによる HPSG 統語解析の評価実験は統語解析技術のさらなる向上のための道筋を示すと考えられる．また実用面では，統語構造が自動的に計算できるようになったことで，様々な自然言語処理への応用が期待される．