様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年５月１２日現在

機関番号：３４４０３
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：２１５００１５２
研究課題名（和文）
コーパスからの語彙項目抽出による大規模な日本語結合範疇文法の構築
研究課題名（英文）
Building a large-scale Japanese Combinatory Categorial Grammar by extracting lexical entries from corpora
研究代表者
大谷　朗（OTANI AKIRA）
大阪学院大学・情報学部・准教授
研究者番号：50283817

研究成果の概要（和文）：本研究では，日本語の実用的な語彙化文法を考察し，コーパスからの語彙項目の（半）自動的な抽出による大規模な文法の構築アルゴリズムを提案した．言語学的な複雑さにより文が長くなる時，パーサーの効率は低下する．そこで，CCG，HPSG，DRT の枠組みに基づいて日本語の複文，複合述語，関係節といった言語学的な問題を分析し，注釈付きコーパスからの CCG 文法の帰納にも利用できる言語学的な形式化に基づいた人間の文処理の方略を示した．

研究成果の概要（英文）： This research project investigated a Japanese practical lexicalized grammar and proposed an algorithm of building a large-scale grammar by (semi-)automatic extraction of lexical entries from corpora. When sentences become longer because of some linguistic complexity the parsing performance deteriorates. Under the framework of CCG, HPSG and DRT, we analyzed such linguistic matters as complex predicates, complex sentences and relative clauses in Japanese, and showed a human sentence processing strategy based on a linguistic formalization, which is also available for inducing a CCG grammar from an annotated corpus.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	1,400,000	420,000	1,820,000
２０１０年度	1,300,000	390,000	1,690,000
２０１１年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：自然言語処理
科研費の分科・細目：情報学・知能情報学
キーワード：CCG・HPSG・DRT・文法・文解析・長文・語彙項目・コーパス

１．研究開始当初の背景

自然言語の計算的性質，特に構造的特性を明らかにしようとする研究には，二つの流れがある．一つは Chomsky の生成文法（1950 年代後半－）であり，もう一つは Ajdukiewicz らの範疇文法（1930 年代後半－）である．

前者は，句構造規則が生成する形式言語の研究に端を発し，情報科学の礎を築いたものの，英語から諸言語へと研究対象を拡げるうちに，生成能力への関心が薄れてしまった．一方，80 年代を中心に，情報科学 にも精通する研究者らは，生成文法の初期の精神と技術を踏襲しつつ，型理論や単一化を導入する

ことで，構文解析器の設計としても通用する厳密な形式化に基づく枠組みを提唱し，今日，文法は理論的な分析だけでなく，工学的にも応用されるに至った．そのような枠組みは，規則を極力廃し，文を構成する重要な情報を語彙に記述することから語彙化文法と呼ばれるが，GPSG，LFG，HPSG，LTAG，そして CCG（結合範疇文法）がこれに属する．また，生成文法では意味の問題を積極的に考えないが，これらの理論では統語構造と意味構造の間に準同型写像を与え，構文解析と同時に構成的意味表示を扱うのも特徴である．

後者は，Frege の関数的解釈意味論を踏襲し，自然言語の表す概念が範疇の階層を持つと考え，言語の構造の中に関数適用，範疇から範疇への関数，関数もまた範疇をなすという性質を観察した．この古典的な範疇文法は，高階型理論の一種とみなすことができ，50 年代後半には Lambek により演繹体系として情報科学と邂逅し，Bar-Hillel らによって文脈自由文法との等価性も示された．Montague の形式意味論の土台に採用されたことでも 60，70 年代には注目を集めるが，50 年代後半に次ぐ範疇文法の転機は，Steedman によって範疇結合規則が見直され，CCG へと拡張された 80 年代後半と言える．

情報科学と言語学における二度の転機を経て，理論的基盤を築いた CCG にとっての，そして言語研究全体にとっての一大転機は，コーパスの利用が大きく進展したことによってもたらされた．90 年代から今日に至るまでに，コーパスに基づく言語処理の研究は，言語研究のほとんどあらゆる分野に浸透し，その重要性が認識されている．そうした情勢の要因の一つは，電子化された大規模な言語データが計算機上で利用可能になったことにある．

米国では，Pennsylvania 大学の Linguistic Data Consortium が，90 年代前半から Wall Street Journal 等のテキストを自動的に解析し，人手で修正することにより品詞情報や構文情報等のタグを付与した，いわゆるタグ付きコーパスを提供しはじめた．これにより，共通のデータを利用して，客観的な研究を行うことが可能となり，Penn Treebank と呼ばれるこのコーパスは，言語処理研究の gold standard として，今日広く用いられている．

そうした応用の一つに，英国 Edinburgh 大学の Steedman が率いる CCG Group が行った 2000 年頃からの一連の研究がある．Clark，Curran，Hockenmaier らは，統計的タグ付けや高被覆な構文解析器の研究等で重要な成果をあげており，特に Hockenmaier は Penn Treebank から CCG Bank（CCG の文法タグ付きコーパス）を作成する等，言語学と計算機科学の手法を緻密に組み合わせた独創的な研究を行っている．

日本語生成文法研究が進展したのは 80 年代中頃からであるが，当時の研究は既に情報科学と乖離していた．日本において語彙化文法に注目したのは工学者であり，それを応用した規則や制約に基づく処理が試みられたものの，80 年代末には限界が感じられていた．

しかし，90 年代半ばになると，新聞記事をデータとして，電子化辞書研究所が EDR コーパスを，京都大学が京都テキストコーパスを公開したことで状況は一変し，言語処理研究は，自然言語の構造的特性とは直接的な関係のない，確率モデルで捉えようとする統計的言語処理として息を吹き返した．

この統計的言語処理は語彙化文法と近接し，2000 年頃からコーパス指向の文法研究が興隆した．国際共同研究組織 DELP-in /HPSG，Parallel Grammar Project/LFG に参加する東京大学，NTT，Fuji Xerox 等が，文法に基づく構文解析器・コーパスの開発，コーパスからの文法抽出を行ってきている．

２．研究の目的

本研究は，語彙化文法の枠組みに基づいて，新聞記事に現れる文のパターンを理論的に分析するとともに，そうして得た形式化を利用して，コーパスから工学的な応用が可能な文法を自動的に抽出することを目的とする．

CCG，HPSG 等の語彙化文法は，理論言語学と計算言語学の知見を柔軟に取り込みうる伝統的かつ最先端の言語理論であるが，CCG は日本語解析への適用があまり試みられていない．そこで，本研究では，理論モデルの設計および大規模データの収集といった両分野が得意とする手法を CCG，HPSG の上で組み合わせ，言語学的に裏付けられた日本語の文法・コーパスといった言語資源の効率的な開発を目指す．

コーパス利用の簡便化に伴い，統計的言語処理が盛んに研究されているが，多くは表層的な処理が対象であり，詳細な意味解析等の深い言語処理を行うには適切でない．また，HPSG，LFG に基づく研究も行われてきているが，それらは文法の形式化を重視した処理の高速化･効率化の研究(HPSG:東京大学の Enju，LiLFeS，RenTAL 等）か，あるいは商業的理由で研究成果の言語資源が公開できない研究（HPSG：NTT の Hinoki，Lexeed，LFG：Fuji Xerox の XLE 等）であり，深い言語処理への応用を指向した文法や，そうした枠組みのもとでタグ付けされたコーパスは公開されていない．そこで，本研究は CCG，HPSG に基づき，統語と意味に関する日本語の構造的特性を明らかにする理論研究を行うとともに，言語学的な偏向を極力排除した工学的な応用への調整が容易な文法を設計する．

そのような精細かつ実用的な文法の開発には，コーパスが不可欠であるが，本研究では，成果が公開されている既存のタグ付きコーパスをもとに，英語 CCG Bank の構築等で確立されている手法を応用し，効率的なタグ変換と文法抽出を行うことで，理論的に設計した文法を最適化し，また実用性の検証の一環として，CCG タグ付きコーパスを試作する．

このように，既存の言語資源の恩恵を最大限に享受する形で研究を推し進め，そうして得た理論的・実証的研究の成果を公開できることを目指す．

３．研究の方法

本研究は，語彙化文法の精細な文法規則・語彙項目を得るため，CCG，HPSG に基づいて言語学的に偏向しない程度に理論的検討を踏まえながら，文法の核となる部分のみを手で記述する．そして，それをテンプレートに解析済みの新聞記事コーパスを入力として，一貫した規則および大量の語彙項目を（半）自動的に抽出することで作業の効率を高め，同時に精度をも高めた工学的に応用可能な実用日本語文法を構築する．

(1) 語彙化文法に基づく統語と意味のインターフェース：日本語長文の要因の再考

新聞記事に，また日本語に限ったことではないが，長文となる要因は以下に大別できる．

i. 複合名詞・複合動詞等の複合語
ii, 重文構造・複文構造
iii. 付加等による修飾

これらは，言語理論的に見ても，また工学的応用を考えても緊要な課題として周知されており，このうち複合動詞と重文に関しては，研究代表者は既に Steedman とともに CCG の枠組みで検討している．

①認識動詞「思う」の表層構成性

新聞記事に現出する複合名詞のほとんどは，例え形態素解析が済んでいても，その多様な構成素関係を理論のみで捉えることは実際上不可能であり，また付加等による修飾については，係り受け解析が済んでいても，項/付加語の判別および修飾関係の意味記述が単純ではないことから，理論的骨子の整備やデータの十分な検討が済んでいない段階で扱う課題としては適切ではない．

そこで，本研究の理論面では，補文構造に関する言語情報の形式化について検討する．一見，言語情報の局所性に反するかのように思われる認識動詞「思う」の統語的・意味的特性を CCG に基づいて厳密に形式化することは，精細な文法を記述する上でも必須である．

②関係節の漸進的解釈のメカニズム

CCG は英語の関係節や等位構造等が簡潔に記述できることから，日本語に適用することで，被覆率の高い文法が作成できると期待されているが，そうした見込みは早急である．これらの言語現象は，実際日英語間では並行していない．例えば日本語には関係詞はなく，また疑似関係節と呼ばれる構文は英語には存在しない．

そこで，本研究は，英語の関係代名詞を含む文に相当する文が，日本語ではどのように具現化し，また解釈されているのかを明らかにし，そうした諸般の性質が説明できるような形式化を検討する．

(2) CCG コーパス試作を指向した語彙項目と文法の抽出

①解析済みコーパスからの語彙項目の抽出

語彙化文法の語彙項目・文法の抽出には，その枠組みに基づいた構文情報タグ付きのコーパスが必要となる．しかしながら，そのような解析済みコーパスは，日本語には存在しないので，本研究では，以下の手順 i-iii により，京都テキストコーパスのタグで表された言語情報を変換することで，CCG の語彙項目・範疇を抽出する．

i. 二分木化：
  a. 文節間の依存構造を二分木に変換
  b. 文節内の形態素リストを左下がりの二分木に変換（暫定的な処置）
ii. 範疇付与：
  a. 述語以外の品詞を CCG の範疇に変換
  b. 部分木の右端に応じた CCG の規則を適用し，両端を支配する節点の CCG の範疇を導出
iii. 範疇抽出：
  右端の述語を支配する節点の左端を走査し，項を補完することで，述語に対する CCG の範疇を抽出

対象言語の性質とコーパスの情報の違いにより詳細は異なるものの，入力文を二分木化し，そこから規則やヒューリスティクス等を用いて CCG の範疇を抽出する手法は，英語 CCG Bank，またそれを参考に行われたドイツ語，トルコ語の実験も基本的には同じである．

小嶋らは，日本語が主要部後置言語であることを利用して，コーパス作成における先の手続き（ib）を簡略化し，効率的に語彙項目を抽出している．しかし，そうした手法にはいくつか問題点があり，語彙項目の抽出率は 6 割程度に留まっている．

そこで，本研究では小嶋らの研究と競合し，同程度の結果を得ることのできる文法を構築することを最初の到達目標とする．そして，次の段階では，（ib）を詳細化することで，さらなる抽出精度の向上を目指す．

②抽出した語彙項目・文法の調整

小嶋らの手法の限界は以下にある.

i. 複合名詞を含む文節の処理

ii. 項/付加語の区別に関して特別な措置がなされていない,

日本語複合語の性質上，主要部後置を利用した構造の簡略化は，動詞の連鎖では有効であるが，名詞相当語の連続では左下がり構造になるとは限らないので，意味構造と対応した精細な構造が抽出できない.

近似解を考えるならば，京都テキストコーパスの下位品詞分類を利用した木構造化を試みる必要がある.（ii）は，元のコーパスの依存解析がそれらを区別していないことに起因すると考えられるが，これも左右の句の主要部同士の関係を調べることで，ある程度は区別できる.

４．研究成果

本研究では,CCGやHPSG等の語彙化文法と，モンタギュー文法や DRT（談話表示理論）等の形式意味論に基づいて，新聞記事に現れる文のパターンを理論的に分析するとともに，そうしたパターンの制約に関連する人間の文理解メカニズムをモデル化し，言語の情報記述の枠組みを精緻化した．また，そうして得られた形式化を利用して，新聞記事データから工学的な応用が可能な文法を自動的に抽出することを試み，さらにモデルの働きをシミュレーションし，言語心理実験の結果と比較することで，提案する言語情報の形式化・モデルの妥当性を高めた.

CCG, HPSG, DRT は，理論言語学と計算言語学の知見を柔軟に取込んだ言語理論であるが，こうした枠組みの日本語解析への適用はあまり試みられていない．そこで，本研究では，理論モデルの設計および複雑な言語情報の形式化，大規模データの収集といった両分野が得意とする手法をこれらの枠組みの上で組合わせ，言語学的に裏付けられた日本語文法・文理解メカニズム・コーパスといった言語情報の形式化，処理方法，資源の総合的かつ効率的な開発を行った.

(1) 新聞記事の文のパターンの一つである重文・複文構造による長文に関し，言語情報処理の形式化における緊要な課題として，以下の二点に問題を絞り，複文構造に関する言語情報の理論的形式化について検討した.

i. 補文の構成要素と考えるべき名詞句が，主文の構成要素であるかのように振舞う統語的問題
（依存関係の交差を許してしまうと，妥当な時間で解析可能な文脈自由規則に文法が収まらない.）

ii. 構文全体の意味が，補文の表す命題に関して閉じていない意味的問題
（構成的意味論の局所性に反してしまうと，統語と意味との間の準同型写像の関係が成立しない.）

一見，言語情報の局所性に反するかのように思われる上記の問題を CCG に基づいて厳密に形式化することは，精細な文法を記述する上でも必須である．本研究では具体例として日本語認識動詞「思う」に関する現象を取り上げて分析し，従来の日本語分析とは異なる，表層構成性に基づく新しい分析を提案した.

(2) 単文において，文を長く複雑にする複合述語に関し，言語情報の局所性に反するかのように思われる構文の多元的制約を厳密に形式化した．特に，補助動詞「-てやる」を含む構文に関する現象を取り上げて分析し，従来の日本語理論分析とは異なる表層構成性に基づく新しい分析を提案した.

(3) 助詞・マーカ分類される範疇について，それぞれ焦点・主題といった日本語の頻出かつ重要な現象に関し，詳細な語彙項目の記述を試み，精緻な理論的説明を提示した.

日本語に限らず，焦点・主題という情報は，自然言語の文集合を結束した一つのテキストとして成り立たせている重要な言語情報である．特にハ・ガの分布，機能に着目し，それらがマークする構成素の文中における談話・意味的役割の説明を，局所的制約の記述として与えたことに意義がある.

(1)-(3)の分析は，従来の計算言語学的アプローチが，日本語の単なる主辞後置性として簡略化，捨象していた言語情報の形式化と異なり，語彙化文法の制約のもと，多元的言語情報の制約として実装に適した見通しのよい形式化を行った点に特徴がある.

(4) 新聞記事の文のパターンの一つである関係節による長文に関し，人間にとっても，また計算機にとっても望ましい解析方法と考えられる漸進的解析に着目し，それを可能とする枠組みを具体的に記述した.

従来の主要な研究対象であった関係節によって引き起こされる構造的曖昧性の解消の問題に焦点をあて，この処理に統語・意味・語用論的制約がどのように用いられるかということを考慮して，DRT に基づいたモデルを構築した．文の持つ具体的な意味表示が完成される過程をシミュレートし，さらにそうした過程が人間の文解析の方法として妥当な振る舞いとなっているかどうかを言語心理学的にも検討した.

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４件）

① Akira Ohtani and Takeo Kurafuji. Quantification and the Garden Path Effect Reduction: The Case of the Universally Quantified Subject. Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25). 査読有. 2011. 41-50.

② Akira Ohtani. Integrating Japanese Particles Function and Information Structure. Computational Linguistics and Intelligent Text Processing: 12th International Conference CICLing 2011, Proceedings, Part 1. Springer Verlag Lecture Notes in Computer Science. 査読有. 6608. 2011. 353-367.

③ Akira Ohtani and Mark Steedman. A Multi-Dimensional Analysis of Japanese Benefactives: The Case of the Yaru-Construction. Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC 24). 査読有. 2010. 503-510.

④ Akira Ohtani and Mark Steedman. Note on Japanese Epistemic Verb Constructions: A Surface-Compositional Analysis. Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 23). 査読有. 2009. 395-404.

〔学会発表〕（計４件）

① Akira Ohtani and Takeo Kurafuji. Quantification and the Garden Path Effect Reduction: The Case of the Universally Quantified Subject. The 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25). December 16, 2011. Nanyang Technological University, Singapore.

② Akira Ohtani. Integrating Japanese Particles Function and Information Structure. The 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2011), February 21, 2011. Waseda University, Tokyo, Japan.

③ Akira Ohtani and Mark Steedman. A Multi-Dimensional Analysis of Japanese Benefactives: The Case of the Yaru-Construction. The 24th Pacific Asia Conference on Language, Information and Computation (PACLIC 24). November 6, 2010. Tohoku University, Sendai, Japan.

④ Akira Ohtani and Mark Steedman. Note on Japanese Epistemic Verb Constructions: A Surface-Compositional Analysis. The 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 23). December 4, 2009. City University Hong Kong.

６．研究組織
(1)研究代表者
大谷　朗（OTANI　AKIRA）
大阪学院大学・情報学部・准教授
研究者番号：50283817

(2)研究分担者

(3)連携研究者