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　２７　年　　６　月　１３　日現在

機関番号：　３４４０３
研究種目：　基盤研究(C)
研究期間：　2012～2014
課題番号：　２４５００１８９

研究課題名（和文）組合せ範疇文法を用いたコーパスからの論理形式の抽出

研究課題名（英文）Extracting logical forms from corpora by using Combinatory Categorial Grammar

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

交付決定額（研究期間全体）：（直接経費）　　4,100,000円

研究成果の概要（和文）：本研究では組合せ範疇文法（CCG）に基づく日本語の語彙化文法を設計し，CCGの派生を利用したコーパスからの論理形式の（半）自動抽出アルゴリズムを考察した．言語学的な複雑さにより文が長くなると解析器の効率は低下する．そこで，CCGの枠組みに基づいて，日本語の多重ガ格，ガーノ交替，場所格ニーデ交替といった言語学的な問題を分析し，タグ付きコーパスからの文法の帰納にも利用できる統語・意味文解析の方略を提案した．

研究成果の概要（英文）：This research project designed a Japanese lexicalized grammar based on Combinatory Categorial Grammar (CCG) and considered an algorithm of (semi-)automatically extracting logical forms from corpora by using CCG derivations. When sentences get longer because of some linguistic complexity the parsing performance deteriorates. Under the framework of CCG, we analyzed such linguistic matters as multiple-GA, GA-NO conversion and locative NI-DE alternation in Japanese, and proposed a syntactic and semantic parsing strategy, which is also available for inducing a CCG grammar from an annotated corpus.

研究分野：　理論言語学・計算言語学

キーワード：　語彙化文法　組合せ範疇文法　CCG　論理形式　言語情報抽出　コーパス　語彙項目　格交替

１．研究開始当初の背景

本研究の代表者（以下，代表者）は 1999 年に語彙化文法理論の一つである主辞駆動句構造文法（Head-driven Phrase Structure Grammar: HPSG）に基づいた日本語の形式文法を記述し，その厳密な形式化がそのまま構文解析器として実装できることを示すとともに，言語の形式理論と実際の言語処理とを結びつける具体的な手法を提案した．この小規模ながらも精細な文法を発表して以来，言語処理研究の一部では，こうした文法に基づく構文解析が見直され，代表者の研究は日本語文法の知見に立脚した精緻な解析を行う日本語句構造文法（Japanese Phrase Structure Grammar: JPSG）として言及されている．しかし，その一方で JPSG のような文法指向の文解析の問題点として，日本語の文法事項を網羅的に含み，高い解析カバー率を実現する文法規則を記述することは容易ではないという指摘もなされていた．

代表者が日本語文法の記述に用いた語彙化文法は，以下の特徴をもつ自然言語の文法理論である．

(1) 文法を個別の規則の集まりとしてみるのではなく，文法規則といえるものは，言語のもつ文構造全般にわたる性質や制約を記述するものとみなす．
(2) 単語（語彙項目）はそれぞれの役割を予測できるほどの内部情報構造をもつ．

JPSG が評価されたのは，日本語に特徴的な言語現象を文が内包する言語情報構造として，日本語の一般的なスキーマ（上述の特徴 1）とそうした文を構成する単語の内部構造が提供する部分的な情報（同特徴 2）に見通しよく振り分けることができたことによる．1 は解析によって生じる規則の組合せ爆発を抑える上でも効果的であったが，小規模な JPSG では形式化していた単語数の少なかったことが，被覆率の問題に関する直接的な原因であることを代表者は認識していた．また，過去の言語学的知見を参考とした言語現象の選定と，それを説明するのに必要な語彙項目の内部構造の設計も十分ではなかった．

こうした課題に対して，語彙項目の量と質に関する二つの方向から，代表者は JPSG の改良を 2006 年頃まで行った．語彙知識ベースの構築では，コーパスから NTT 日本語 HPSG の語彙項目約 3 万 2 千語を抽出し，実世界テキストの解析に基づく知識獲得の実験を行った．また，JPSG の語彙項目に対する意味・談話情報の拡張，そうした言語情報構造のタイプ階層化等，情報学的観点からさまざまな見直しを行うことで，言語学的な分析対象が拡大可能となった NAIST JPSG を実装した．

そして，このうちの語彙項目の内部構造に関する研究が評価され，2007 年以降は，意味･談話処理に関する HPSG の語彙記述の研究に加えて，Mark Steedman 教授の指導の下で教授が提唱するもう一つの語彙化文法理論である組合せ範疇文法（Combinatory Categorial Grammar: CCG）に基づく日本語形式文法に関する研究も行ってきていた．しかしながら，日本語 CCG の研究に関しては，語彙項目の量に関する取組みが十分には行えておらず，また，同じ語彙化文法ではあるものの，HPSG と CCG では語彙項目の具体的な記述に際して地道に行なっていかなければならない言語現象の形式化における文法観や分析が異なっているため，コーパスからの語彙項目の抽出に関しても検討することが多い．

そこで，本研究は CCG に基づく言語情報の形式化を検討し，コーパスからの言語情報抽出の応用に展開するための基礎研究を行う．計画をすすめていく上で，代表者は以下の知見，予備的な研究成果を得ていた．

① HPSG と同様に CCG に基づく言語情報の形式化では，自然言語の特徴的な現象に関しても，基本的には意味・談話情報の表示は構文構造の構成性に対応する．
② Steedman による範疇結合規則を導入した CCG では，構文構造は意味表示を出力するためだけに使われ，自然言語の構成性に合わない構文解析を行ったとしても，同じ論理形式が得られる．
③ ②により，HPSG よりも CCG に基づく構文解析器の方が頑健になる可能性がある．
④ 上記①，②により，CCG は日本語に特徴的な現象に関しても，理論的な説明が可能である（例えば，複合述語構文，補文構造，焦点・主題化，助詞など）．
⑤ 語彙化文法理論は構文構造と意味構造の間に準同型写像を与える．この特徴により，コーパスに構文情報をタグ付けして，文法規則を逆適用すると，述語の語彙項目が獲得できる．

２．研究の目的

本研究は，文の構造と意味との関係を明示的に理論の中に取り込んだ CCG に基づいて，形態・構文情報が付与されたテキストコーパスから意味情報を（半）自動的に抽出することを目的とする．この目的を達成するために，本研究では段階的かつ具体的な以下(1)~(3)の詳細な下位目標を設定する．

(1) 構文構造とその論理形式を CCG に基づき一般的に記述する方法を提案し，こうした CCG の記述を構文情報が付与されたテキストからの意味情報の抽出に使用する方法を検討する．
(2) より汎用的な抽出を目指して，構文構造を論理形式に変換するプログラムの基本試作を行う．
(3) プログラムをコーパスに適用する実験を行い，適切な論理形式を多く抽出する方法を考察する．

３．研究の方法

上記の研究の背景およびそれを踏まえた具体的な目標を遂行するために，本研究はCCG に基づく語彙項目の具体的な記述方法を確立し，意味情報の抽出を行うための構文構造から論理形式への変換やコーパスからの論理形式の(半)自動抽出への応用に展開するための基礎研究を行う．そして，研究期間内には，分析や実験を行い，以下のことを明らかにする．

① 日本語に特徴的な言語現象を選定，追加し，CCG に基づく言語学的な分析，記述を行う（量化に関する研究を行ってきているので，否定や作用域に関する意味論的問題も扱う)．
② 構文構造とその論理形式を蓄積し，対応付けの一般的な制約を CCG に基づいて形式化する．
③ ②の制約を用いて，構文情報が付与された文から論理形式を抽出するアルゴリズムを設計する．
④ より汎用的な抽出を目指して，構文構造を論理形式に変換するプログラムの基本試作を行う．
⑤ プログラムを構文構造タグ付きコーパスに適用して，論理形式を大量に抽出する実験を行う（既存のタグ付きコーパスの構文構造をそのまま利用した論理形式の抽出が困難である場合は，先にコーパスの構造変換を行っておく)．
⑥ 適切な論理形式が抽出できない文を精査し，可能ならばCCG に基づく分析を与える（⑤での語彙項目抽出のために行った構文解析のカバー率 60%弱をひとまずの目標とする)．
⑦ ④のプログラムを改良し，再度実験およびその評価を行い，研究を総括する．

(1) CCG に基づく日本語の言語学的分析

代表者は既にいくつかの日本語に特徴的な言語現象に関する分析を行ってきていたが，上述の課題(1)では，それら以外の現象に対してCCG に基づく言語学的な分析を行う．

① 日本語に特徴的な言語現象の選定と CCG に基づく形式化

一般に文は長くなればなるほど解析は困難になる．コーパスに数多く含まれる新聞記事に，また日本語に限ったことではないが，文が長くなる要因は，次の三つに大別できる．

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

(i)-(iii)は，言語学的に見ても，また言語処理への応用を考えても緊要な課題である．このうち(i)の複合動詞と(ii)の重文に関しては，代表者は既にいくつかの現象を CCG の枠組みで検討しているが，個別の現象を扱ったのみで体系的ではない．(i)の複合名詞と(iii)の付加の問題については課題(2)-③で述べ，ここでは複合動詞・重文に関し，以下の二点に問題を絞り，補文構造の形式化について検討する．

a. 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)
b. 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，統語と意味の間の準同型写像関係が成立しない)

一見，言語情報の局所性に反するように思われる a，b を CCG に基づいて形式化することは，精細な文法を記述する上でも必須であり，本研究では特に繰り上げ構文を分析する．

(2) 構文情報から意味情報への変換

構文構造と意味表示の間の準同型写像関係と，課題(1)を行うことで得たそれらの対応付けに関する CCG に基づく制約を用いて，構文情報が付与された文から論理形式を抽出するプログラムを試作する．

② CCG に基づく構文構造と論理形式の対応付けの一般化

構文構造とその論理形式の対応に関する一般的な制約を CCG に基づいて形式化する．

③ 構文情報が付与された文から論理形式を抽出するアルゴリズムの設計

既存のタグ付きコーパスの構文構造の仕様に従いタグ付けした文をもとに考察する．

付加等による修飾(課題(1)-①より)については，係り受け等の解析が済んでいても，項と付加語の判別が単純ではなく，適切に処理するアイデアを持ち合わせていない．

CCG の文法（語彙項目，論理形式）を抽出するには，CCG に基づく構文情報がタグ付けされた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ないので，ここでは以下の手順(i)-(iv)により，京都テキストコーパスのタグを変換することで，必要なコーパスを作成する．

i. 前処理： 複合名詞（課題(1)-①）のまとめあげ
ii. 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二分木化．
iii. 範疇付与：a. 述語以外の品詞を CCG 範疇に変換，
　　b. 部分木の右端に応じた CCG 規則を適用し，両端を支配する節点の CCG 範疇を導出．
iv. 範疇抽出： 右端述語を支配する節点の左端の走査，項の補完により述語の CCG 範疇を抽出．

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文法理論のみで捉えることは実際上不可能なので，(i)は一語としてまとめ上げる．

対象言語の性質と解析済みコーパスの情報の違いを反映して詳細は異なるが，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，Steedmanの研究グループの手法と基本的には同じである．

④ 構文構造を論理形式に変換するプログラムの基本試作

より汎用的な抽出を目指して，構文構造を論理形式に変換するプログラムの基本試作を行う．

(3) コーパスからの論理形式の抽出

課題(2)で試作したプログラムを適当なコーパスに適用して，論理形式を抽出する実験と評価を行う．

⑤ 実験，評価

プログラムを構文構造タグ付きコーパスに適用して，論理形式を大量に抽出する実験を行う．

⑥ 問題となる文の精査

適切な論理形式が抽出できない文を精査し，可能ならばCCGに基づく分析を与える(語彙項目抽出のために行った予備的実験で得た構文解析のカバー率60%弱を当面の目標とする)．

⑦ プログラムの改良，研究の総括

プログラムを改良し，再度実験およびその評価を行い，研究を総括する

４．研究成果

本研究では，研究方法(1)～(3)に示した課題に関して段階を追って遂行した．

(1) CCGに基づく日本語の言語学的分析
（第Ⅰ期：平成24年度～平成25年度）
(2) 構文情報から意味情報への変換
（第II期：平成25年度後半～平成26年度）
(3) コーパスからの論理形式の抽出
（第III期：平成26年度後半）

(1) CCGに基づく日本語の言語学的分析

代表者が行った語彙化文法に基づく語彙項目の内部構造に関する一連の研究（NAIST JPSG, Steedmanとの共同研究）は，日本語の主要な特徴的現象を概観しているが，説明可能な現象が増えるほど文法は精緻になる．

そこで，まず第Ⅰ期の課題として，未整理の特徴的な現象から主要なものを選定し，CCGに基づいて形式化を行った．

（こうした作業は，言語学的に新規の分析を提案するだけでなく，第II，III期のプログラムによる解析の被覆率を上げることにもつながる．）

［雑誌論文⑤，学会発表④］

特に，代表者が理論的に整理した現象のうち特徴的なものは，日本語におけるガ格句の分布である．日本語のガ格は主語をマークする後置詞として文中に一つ含まれるのが基本的な用法であるが，一部の動詞がガ格目的語を選択することに加え，文頭に多重なガ格（あるいは主語）が生起することから，そうした文は長くなる傾向にある．また，そのようなガ格の一部はノ格に置き換えが可能であり，いわゆるガ—ノ交替という現象も起こることから，複数のガ格を含む文を精細に解析することは概して困難であった．

本研究ではこうした多重ガ格構文を整理し，その種々の現象の説明に最低限必要とされる後置詞ガ・ノに関する複数の語彙項目をCCGに基づいて記述し，文解析にも適用可能な言語学的な分析を提案した．

(2) 構文情報から意味情報への変換

代表者はこれまで日本語に特徴的な現象を体系的に説明するために，語彙化文法理論に基づいて，語彙項目の内部情報構造に関する研究を行ってきたが，特に本研究では文の構造と意味との関係を明示的に文法理論の中に取り込むことで，これまで手動で記述していた意味情報を，構文情報が付与されたテキストコーパスから（半）自動的に抽出するために必要な言語分析を行った．

意味論・意味情報処理に関する問題は枚挙に暇がないといってよいほどだが，そのうち既存の言語理論で形式化が可能なものとなると問題は限定される．本研究では文処理における量化の効果および語彙意味の諸問題を検討した．

［雑誌論文④］

過年度の心理言語学的な実験では，普遍量化された名詞句「すべての NP が」の解釈が解析の過程において一時的に曖昧な状態に陥る文が，量化子を伴わない裸名詞句を含む文に比べ，ガーデンパス文として解釈される効果が減少した．このことは談話表示構造(Discourse Representation Structure: DRS)が漸進的に構築されると仮定することで，以下のように説明される．

文中の名詞句に普遍量化表現が含まれる場合は，その意味構造に関連する二つDRSが導入されるため，一時的な曖昧性に関しては解釈の余地が生じる．一方，裸名詞句の場合には，単一のDRSだけが導入されことになり，唯一の解釈が得られる．その結果，後続する名詞句が読み込まれると，必然的に再解釈が必要となり，ガーデンパス効果が生じる．

［雑誌論文②，③，学会発表②，③］
CCG は他の文法理論では形式化が困難な長距離依存や並列構造における述語の項構造の同定が可能であるものの，格や項の交替において対応する項を同定する機構，いわゆるリンキングを備えていない．そこで，本研究では日本語の場所格交替を取り上げ，概念意味論 (Conceptual Semantics: CS)のリンキング機構を CCG に取込む拡張を試みた．

述語論理を基本とする CCG の意味論に CS を導入すると，場所を表す後置詞ニとデの分布および格交替の説明が可能となる．この研究では二つの統語・意味一致規則を導入し，動詞の意味クラスによって異なる語彙概念構造がこれらの規則と作用し，BE 関数を含む動詞はニ格要素を伴うが，それを含まない動詞はデ格要素を伴い，また交替が起こる動詞は多義であることを示した．

(3) コーパスからの論理形式の抽出

前段階で遂行してきた日本語に特徴的な言語現象の CCG に基づく形式化を踏まえて，本研究の最終段階では構文構造から意味構造への変換アルゴリズムを設計し，実装した変換プログラムにより構文情報付きコーパスから論理形式を(半)自動的に抽出することを目指した．適切な論理形式が得られない文については個別に精査し，可能であれば CCG に基づく分析を与え，実験と評価を経て研究を総括した．

［雑誌論文①，⑥，学会発表①，⑤］
コーパスを利用した本研究における定期的かつ継続的なアウトプットとして，また，言語情報の抽出に関する基礎的な研究およびその具体的応用を模索するチャレンジ的な課題として，研究の各段階で蓄積したノウハウを活かしつつ，小規模なコーパスの調査でも効果が得られる学習者コーパスの研究も付随的に行った．

スペイン語ネイティブ教師の助力を得て，日本人スペイン語学習者の作文を収集して小規模な学習者(エラー)コーパスを作成し，そこに含まれる文法誤り，特に性・数の一致に関する誤り（雑誌論文⑥，学会発表⑤）および冠詞の定・不定・無に関する誤り（雑誌論文①，学会発表①）を自動的に検出する実験を行ったが，学習者エラーコーパスを形態素解析した出力に関して Karlsson の制約文法 (Constraint Grammar: CG) に基づいて記述した文法を用いて解析したところ，想定した文法誤りを多数検出することができた．

また，こうした実験において検出が失敗した例を精査してみると，複数作業者によるタグ付与作業の精度が均一でないことに起因するものの占める割合が無視できないほどあった．そこで，実験的な小規模コーパスを今後より大規模なものへと発展させるための準備として，複数作業者によるタグ付与を統制する方法についても検討した．

(1)～(3)の総括

近年のコーパス利用の簡便化と相俟って，CCGやHPSGといった語彙化文法理論を基盤とした研究は，世界中で盛んに行われている．日本もこうした理論に基づく大規模文法の記述に取り組める状態になってきたが，そのような研究のほとんどは，意味表現の出力を伴わない構文解析器の高速化・効率化を指向したものであった．

代表者は，文法指向の文解析が日本でも盛んになる前より，HPSG に基づく言語情報の形式化の研究を展開し，近年は CCG に基づく語彙項目の内部構造の記述に関し，国外の研究者と共同研究を行ってきた．今日では代表者以外の研究者も，研究の詳細は違えども CCG などの語彙化文法に基づいて，言語情報の基礎・応用研究を国内で進めてきている．

しかしながら，そうした状況であっても，日本語の言語分析や処理では多くの課題が残されている．本研究は，そうした課題に取組み，CCG に基づく日本語の言語研究を英語に関して行なわれている研究と同等のレベルに引き上げるための基盤作りを試みた．

代表者が共同研究を行っているエジンバラ大学 Steedman 教授のグループは，教授の言語学的な基礎研究に加え，共同研究者らが高い被覆率の解析器の実装，統計情報に基づくコーパスの自動タグ付与，ツリーバンクからの CCG タグ付きコーパスへの変換といった情報学的な応用面でトップレベルの成果をあげており，語彙化文法理論の研究をリードしている．それに対して代表者の研究は，個人が推進する研究ゆえに量の面では比肩できなかったものの，いくつかの話題においては質の面で同等であった．そうしたことから本研究の目標「言語学的な基盤研究から情報学的な応用までの CCG に基づく一貫した日本語研究の基盤の確立」は概ね達成できたと思われる．

５．主な発表論文等

〔雑誌論文〕（計６件）

① Maria del Pilar Valverde Ibanez and Akira Ohtani, Annotating Article Errors in Spanish Learners Texts: Design and Evaluation of an Annotation Scheme, Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation (PACLIC 28), 査読あり, 2014, pp.234-243.

② Akira Ohtani, Locative Postpositions and Conceptual Structure in Japanese, Proceedings of the 27th Pacific Asia Conference on Language, Information and Computation (PACLIC 27), 査読あり, 2013, pp.382-389.

③ Akira Ohtani, Locative Alternation and Conceptual Structure in Japanese: A CCG Approach, Abstract Book of 9th International Conference on Cognitive Science (ICCS2003), 査読あり, 2013, p.7.

④ Akira Ohtani, Takeo Kurafuji and Masakatsu Inoue, Quantification and the Garden Path Effect Reduction: The Case of Universally Quantified NP, International Journal of Asian Language Processing, 査読あり, Vol.22, No.3, 2013, pp.127-146.

⑤ Akira Ohtani and Maria del Pilar Valverde Ibanez, Nominative-marked Phrases in Japanese Tough Constructions, Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation (PACLIC 26), 査読あり, 2012, pp.272-279.

⑥ Maria del Pilar Valverde Ibanez and Akira Ohtani, Automatic Detection of Gender and Number Agreement Errors in Spanish Texts Written by Japanese Learners, Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation (PACLIC 26), 査読あり, 2012, pp.299-307.

〔学会発表〕(計５件)

① Maria del Pilar Valverde Ibanez and Akira Ohtani, Annotating Article Errors in Spanish Learners Texts: Design and Evaluation of an Annotation Scheme, The 28th Pacific Asia Conference on Language, Information and Computation (PACLIC 28), December 12, 2014, Phuket, Thailand.

② Akira Ohtani, Locative Postpositions and Conceptual Structure in Japanese, The 27th Pacific Asia Conference on Language, Information and Computation (PACLIC 27), November 22, 2013, Taipei, Taiwan.

③ Akira Ohtani, Locative Alternation and Conceptual Structure in Japanese: A CCG Approach, 9th International Conference on Cognitive Science (ICCS2003), August 27, 2013, Kuching, Sarawak Malaysia.

④ Akira Ohtani and Maria del Pilar Valverde Ibanez, Nominative-marked Phrases in Japanese Tough Constructions, The 26th Pacific Asia Conference on Language, Information and Computation (PACLIC 26), November 8, 2012, Bali, Indonesia.

⑤ Maria del Pilar Valverde Ibanez and Akira Ohtani, Automatic Detection of Gender and Number Agreement Errors in Spanish Texts Written by Japanese Learners, The 26th Pacific Asia Conference on Language, Information and Computation (PACLIC 26), November 8, 2012, Bali, Indonesia.

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