様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １８ 日現在

機関番号：３４４０３
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１２１０８

研究課題名（和文）ツリーバンク主導で構築する日本語主辞駆動句構造文法の形式化に関する研究

研究課題名（英文）A treebank based formalization of a Japanese Head-driven Phrase Structure Grammar

研究代表者
大谷　朗（OTANI, Akira）

大阪学院大学・情報学部・教授

研究者番号：５０２８３８１７

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では語彙レベルの品詞タグと句レベルの機能情報に関する注釈が付与されたKeyakiコーパスを基にHPSGツリーバンクを構築する方法について考察した。Keyakiの注釈付与スキームは特定の言語理論に拠らないが、我々のツリーバンクは主辞駆動句構造文法(HPSG)に基づくことで統語情報だけでなく他の言語情報を付与することができる。こうしたツリーバンクはHPSGパーサーが解析し、言語学者が確認，修正した句構造木から半自動的に作成することができる。

研究成果の学術的意義や社会的意義
言語学的な分析に基づいて形式化された文法とコンピュータによって実世界のテキストを大量に解析するために設計されたデータ構造とを結び付けようとする本研究は、精細な統語構造と簡単な意味情報の注釈を各文に付与した日本語ツリーバンクを半自動的に構築した。そして、それがより精細なものとなるように改良を重ねていくことで、言語研究の発展だけでなく情報化社会を支える日本語処理の精度向上にとっても有用な言語資源の開発を目指した。

研究成果の概要（英文）：This research project considered a way of building an HPSG treebank on top of the Keyaki parsed corpus, which has already been annotated with part-of-speech tags at the word-level and enhanced with functional information at the phrase-level. While the Keyaki annotation scheme is theory-neutral, our treebank can be annotated with syntactic and other linguistic information following the Head-driven Phrase Structure Grammar (HPSG). The treebank can be created semi-automatically, where an HPSG parser assigns some phrase structure which linguists then check and, if necessary, correct.

研究分野：自然言語処理

キーワード：コーパス　ツリーバンク　HPSG　主辞駆動句構造文法　日本語　言語処理　言語資源

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1) 言語処理技術と親和性が高い HPSG でも大規模な文法の記述は困難.

語彙化文法の一つ主辞駆動句構造文法 (Head-driven Phrase Structure Grammar: HPSG) (①, ②)は，少数の文法原理によって自然言語の規則性を，豊かな語彙項目によって個別言語に特有な現象を説明する言語理論である．素性構造に基づいた厳密な形式化が言語情報に与えられるため，言語処理技術との親和性が高く，応募者は日本語に特徴的な現象をこの枠組に従って記述することで，小規模な文法の解析器（③）を作成している．複雑な文法を効率的に記述するツール LKB (Linguistic Knowledge Builder) (④) なども開発されているが，それでも実世界のテキストを網羅的に解析できる大規模な文法の構築は難しかった.

(2) 文法の自動抽出に適用可能な日本語ツリーバンクがない.

大規模な文法を人間が直接定義するのは困難であるという認識が広まり，また，構文木の定量的な評価も必要であることから，次第に文法はツリーバンクから導出されるものと言う見方が，特に工学者の間で広まった．HPSG も例外ではなく，やはり工学者らが中心となって構文木を付与した最初の大規模な英語ツリーバンク Penn Treebank からの文法の抽出が試みられ（⑤), 応募者も NTT との共同研究（⑥)において，大規模な日本語 HPSG 文法 Jacy（⑦）を拡張し，日本語ツリーバンク Hinoki (⑧) を開発している．ただ，こうしたツリーバンクの開発には莫大な人的・金銭的コストがかかっており，Hinoki の場合は配布禁止のデータも使用されていた．そのため普及することなく，これらの研究は停滞してしまった（パブリック・ドメインデータの使用により 2018 年に Jacy はオープンソース化されたが（⑨)，文法形式化の進展はなかった).

(3) クリエイティブ・コモンズ・ライセンスの Keyaki ツリーバンクがリリース.

2016 年，国立国語研究所は現代日本語の書き言葉テキストに文の統語・意味解析情報のタグ付与を行った NINJAL Parsed Corpus for Modern Japanese (NPCMJ) (⑩) の構築を開始し，現在その一部はウェブ上で公開されている．そして，この NPCMJ の前身である Keyaki ツリーバンク（⑪）がクリエイティブ・コモンズ・ライセンスで利用できるようになり，その潜在的な規則性，すなわち文法の抽出を試みることができる状況となった.

２．研究の目的

(1) 研究課題の核心をなす問い

言語解析に有用な知見は，浅い解析に基づく解析木などの「量」を重視するコーパスからではなく，言語学的に意味のある一般化 (linguistically significant generalization) に基づくことでタグの情報の「質」を重視したコーパスからこそ，精細な文法として抽出できるのではないか.

(2) 「問い」に答えるために

機械可読なデータを集めることが困難であった頃から量が重視されてきたコーパスは，形態素解析の結果をタグ付けした程度のものであっても，タスクによってはそれが解決にもたらす恩恵は計り知れない．しかしその反面，本来は関連付けて考えるべき問題を細分化し，それぞれのサブタスクに必要なコーパスや言語モデルを用意する必要が生じてしまう（例えば，述語項構造解析器 SynCha の省略検出では，省略検出モデル，文内先行詞同定モデル，文間先行詞同定モデルなどの出力結果を整数計画法に基づいて最適化することにより最終的な省略検出の結果を得ている).

しかしながら，Keyaki ツリーバンクを利用すれば，同じタスクが句構造解析のサブタスクとして捉えられ，照応先の同定を伴うことなく省略検出に取り組むことができる．このことはタグ情報に基づく解析木を処理しながら，文法理論としての空範疇原理に関する処理を行っているものとみなすことができる.

(3) 何をどのように明らかにするのか

上記の「問い」に答えるべく，本研究は言語学的な分析に基づいて統語・意味情報がタグ付けされたツリーバンクが内包する言語情報の統括的な形式化として日本語 HPSG 文法を構築する.

３．研究の方法

本研究は Hinoki の開発手法を用いて，ツリーバンクの解析で必要となった文法修正の結果を適用しながら解析を漸進的にすすめていくことで，質と量を兼ね備えた言語資源を開発する．

(1) 精細な文法の構築
・精細な言語情報がタグ付けされたツリーバンクからの文法抽出およびその技法の確立．
・(Jacy に対する) Yet Another Japanese HPSG の構築．

非商用であり，統語・意味解析情報のタグ付けの精度が高い Keyaki ツリーバンクを利用して，言語学的な分析を計算機可読な文法として実用に結び付ける際に生じる諸問題を解決していくことで，言語学的に意味のあるタグ情報から下記 (2), (3) にあげる特徴を持つ文法を抽出する．
そして，長らく一択の状況が続いていた大規模日本語 HPSG において，Jacy に競合する文法を公開することで，大きな進捗が見られなかった文法解析の実用化を推進することを目指す．

(2) 学術的独自性
・深い解析情報を含意するタグ付けに対する適切な文法形式化．
・空の要素の検出など言語現象に関連したタスクへの取組み．

Keyaki が基づく言語分析は，特定の言語理論に拠らず中立である．そのためツリーバンクに付与された解析木の情報と HPSG に基づく木の解析に用いられた理論的デバイスを関連付けるには，原理やスキーマなどの全体的な調整も必要となる．
省略が頻繁に起こる日本語ではその検出が緊要の課題であるが，現状では実用に足るだけの性能に達していない．本研究では空範疇としてタグ付けされた要素に関する統語・意味情報などの生起条件と HPSG の原理やスキーマとの対応関係を精緻化することでこのタスクに取り組む．

(3) 創造・拡張性
・浅い解析がもたらす量だけでなく，深い解析相当のタグ情報を適切に扱い質と量を兼備．
・ツリーバンクに含まれていない文も解析することで，データと文法の両方を拡張．

文法の性能の評価には交差検定を行うか，訓練データとしなかった文をデータ・セットとしてホールド・アウト法を実施する．また，ツリーバンクに含まれていない文についても実験を行う．
文法の構築とツリーバンクの開発は密接な関係にあり，解析結果はデータの拡張と解析器へのフィードバックの両者に利用される．

４．研究成果

本研究は Keyaki ツリーバンク主導で日本語 HPSG 文法を構築した．主要原理やスキーマを適宜調整し，必要に応じて新たな言語制約を導入することで言語解析のための実用的な文法を形式化し，またそうして得た文法に対して評価を行った．

(1) 何をどのように，どこまで明らかにしたのか

以下の段階的，循環的な計画に従い，文法の理論的形式化や構築，検証作業をすすめた．
・ツリーバンクからの文法情報の抽出 → Yet Another Japanese HPSG の形式化（G1）
・文法 G1 の拡張 → 文法 G1 およびその修正・拡張結果を適用したツリーバンキング（G2）
・文法 G2 の評価 → 文法 G2 の修正（G3）

当初の計画から遅れたことは否めず，研究期間内に到達したのは上記の文法 G2 の段階であり，G２ → G3 → G2 … のように文法開発を循環させるまでには至らなかった．

(2) 開発環境基盤の整備と活用
・DELPH-IN ツールの整備（一部のツールは安定して稼働させることができなかった）．
・NPCMJ 検索ツールの整備，活用．

経年により文法開発に必要ないくつかのツールは開発当初とは稼働する計算機状況が大きく異なってきているため，同じ動作環境を構築することができず，特に DELPH-IN ツール（LKB, PET, [incr tsdb()]）の導入には想定以上の時間を要し，終始安定動作を得ることができなかった．
一方，Keyaki，NPCMJ の検索ツールはリリース年が新しいため導入，運用には全く問題はなく，文法開発ツールの導入に試行錯誤しつつ，この検索ツールを活用することでツリーバンクのアノテーション・スキームの分析を行った．

(3) ツリーバンクからの文法抽出
・短文分析の主要パターンの検討
・重文・複文分析の主要パターンの検討

「基礎日本語文法－改訂版－」の各文法項目に該当する用例を中心に，Keyaki における基本構文のアノテーション情報と HPSG 文法の原理的説明との対応関係を精査した．また，Hinoki ツリーバンクの解析に用いた Jacy 文法の形式化を確認し，Keyaki を解析する大規模な HPSG 文法の主要部分を構築し，短文から長文レベルに適応範囲を広げられるように文法を拡張した．

英語やドイツ語などを対象に行われている同様の研究から，後れを取っている点は否めない．主要な文法開発ツールも日本語の解析を念頭に開発されたものはない．しかしながら，近年急速に整備されてきているツリーバンクなどのデータの充実ぶりは海外の研究にも決して劣らない．
本研究において環境整備の困難以外に進捗の妨げになったのは複合や連接といった言語現象への対応である．膠着語でもある日本語に特有なこれらの現象の分析は，日本語研究はもとより言語処理研究においてもツリーバンクの構築に必要な文法として形式化できるほどの包括的な知見はなく，ゆえに新たに分析を進めたものの期間内に結果をまとめるには至らなかった．

＜引用文献＞

① *Head-driven Phrase Structure Grammar*. Pollard, Carl and Ivan A. Sag. Chicago: The University of Chicago Press. 1994.
② *Syntactic Theory: A Formal Introduction*. Sag, Ivan A., Thomas Wasow, and Emily M. Bender. Stanford, CA: CSLI Publications. 2003.
③ *A Constraint-based Grammar Approach to Japanese Sentence Processing: Designing a Systematic Parser for Fundamental Grammatical Constructions and Its Extensions with Semantic and Pragmatic Constraint*. Akira Ohtani. Ph.D. thesis. Nara Institute of Science and Technology. 2005.
④ *Implementing Typed Feature Structure Grammars*. Ann Copestake. Stanford, CA: CSLI Publications. 2002.
⑤ *From Linguistic Theory to Syntactic Analysis: Corpus-Oriented Grammar Development and Feature Forest Model*. Yusuke Miyao. Ph.D. thesis. University of Tokyo. 2006.
⑥ 共同研究. 言語理解のための日本語文法の構築. プロジェクト・リーダー Francis Bond. NTT コミュニケーション科学基礎研究所. 2003-2004.
⑦ *Jacy: An Implemented Grammar of Japanese*. Melanie Siegel, Emily M. Bender, and Francis Bond. Stanford, CA: CSLI Publications. 2016.
⑧ The Hinoki Treebank for Text Understanding. Francis Bond et el. *Proceedings of the 1st International Conference on Natural Language Processing (IJCNLP-04) revised selected papers*. Springer Verlag Lecture Notes in Computer Science. 3248:158-167. 2004
⑨ Introduction and Demo of Jacy: An Implemented HPSG Grammar of Japanese. Moeljadi, David and Takayuki Kuribayashi. Paper presented at the Workshop on the Clause Structure of Japanese and Korean in the 25th International Conference on Head-driven Phrase Structure Grammar (HPSG 2018). University of Tokyo. 2018
⑩ NINJAL Parsed Corpus of Modern Japanese. Version 1.0. National Institute for Japanese Language and Linguistics. 2018.
(https://npcmj.ninjal.ac.jp/interfaces/ accessed 2018/03/13).
⑪ The Keyaki Treebank Parsed Corpus. Version 1.1. Alastair Butler, Kei Yoshimoto, Shota Hiyama, Stephen Wright Horn, Iku Nagasaki, and Ai Kubota. 2018.
(http://www.compling.jp/keyaki/ accessed 2018/03/13).

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関