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　３ 年　 ５ 月 ２５ 日現在

機関番号：６２６１８
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ００５２３

研究課題名（和文）汎用的な範疇文法ツリーバンクの構築

研究課題名（英文）Development of a multi-purpose categorial grammar treebank

研究代表者
窪田　悠介（Kubota, Yusuke）

大学共同利用機関法人人間文化研究機構国立国語研究所・理論・対照研究領域・准教授

研究者番号：６０７４５１４９

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：範疇文法のツリーバンクであるABCツリーバンクを構築し、理論言語学・自然言語処理の学際的な研究のための言語資源としてオープン・アクセスで公開した。本ツリーバンクは、特に、使役や受身などの文末述語の扱いや名詞修飾の構造に関して、構成的意味論を導くために妥当な分析を行っている点に特色がある。また、構築の際に用いたスクリプト、depccgパーザをツリーバンクで訓練したモデルなどの関連ツールもオープン・アクセスで公開した。また、depccgパーザとccg2lambdaのツールキットを使って意味表示を導出するパイロット実験を行い、自然言語処理分野での意味解析の枠組みとしての利用が可能なことを確認した。

研究成果の学術的意義や社会的意義
ABCツリーバンクは、日本語の範疇文法ツリーバンクで公開されているものとして現在唯一のものであり、理論言語学、自然言語処理両分野を取り結ぶ学際的な研究領域の活性化に寄与することが期待される。特に、パーザやツリーバンクなど、自然言語処理のツールを援用した計算論的モデリングに基づく言語研究が近年理論言語学研究において活発になってきているが、計算言語学の究極の課題である、意味解析・意味理解にまで踏み込んだ研究はまだ極めて少ない。範疇文法に基づくABCツリーバンクは、意味解析・意味理解の計算論的研究のための基盤資源としての利用が特に期待できる。

研究成果の概要（英文）：In this project, we constructed a categorial grammar treebank called the ABC Treebank. The ABC Treebank has been made available as an open-access linguistic resource for interdisciplinary research in theoretical linguistics and natural language processing. A characteristic feature of this treebank is linguistically adequate analyses of sentence-final predicates such as causatives and passives as well as the structure of noun modifications. We have released related tools such as the script used for treebank conversion and the depccg parser training model using the treebank. In addition, a pilot experiment was conducted to derive semantic representations using the depccg parser and the ccg2lambda toolkit, and it was confirmed that the treebank can be used as a framework for semantic analysis in NLP research.

研究分野：言語学(統語論、意味論)

キーワード：ツリーバンク　カテゴリ文法　意味解析　日本語　NPCMJ　ABCツリーバンク

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

研究開始当初は、いわゆる AI ブームが言語学の内部でも認知されるようになってきており、深層学習などの機械学習の手法の言語研究への援用なども海外では試みられるようになってきていた。この動きと連動して、国内外で、過去 20 年ほど下火になっていた論理学的な手法を用いた計算言語学研究が再活性化する兆しが見えている状況だった。言語学的に妥当だが計算量の問題から実装が困難な文法理論を計算機に実装して理論の検証を行う試みは、80 年代には先駆的すぎて挫折した。AI 研究など隣接分野において深層学習などの新技術が急速に発展することで、このような研究に再び本格的に取り組む道筋が見え始めてきた。

そのような中で、お茶大(「知識に基づく構造的言語処理の確立と知識インフラの構築」) や国語研 (「統語・意味解析コーパスの開発と研究 (NPCMJ)」)などで言語学の知識に基づくタイプの計算言語学と密接に関連する大規模なプロジェクトが並行して動いており、計算機による自然言語の意味の解析の研究の基礎資源となる、質の高いリソースの開発が急務との認識にいたり、本研究に着手した。

２．研究の目的

上記の動向に鑑み、本研究では、次世代の学際的理論・計算言語学研究の基盤となる言語資源として、汎用的な範疇文法のツリーバンクである ABC ツリーバンクを構築した。この際、範疇文法にはいろいろな流派が乱立しているので、汎用性を高めるため、なるべく特定の枠組みによらないアノテーションを行う方法を考案した。これが、ABC ツリーバンクの背後にある「理論」である ABC 文法である。ABC 文法は、関数適用の規則のみからなる最も単純な範疇文法である AB 文法に関数合成 (Composition) の規則を加えた、アノテーションの便宜のための枠組みであり、CCG やタイプ論理文法などの、特定の範疇文法の理論のいわば「中間言語」とでも呼べるものとなっている。ABC 文法の範囲内でアノテーションを行うことによって、構築したツリーバンクを複数の範疇文法の枠組みに合わせて再変換することが容易となる。このことにより、様々な目的に合わせて利用できる言語資源となる。

ツリーバンクの構築には、「けやきツリーバンク」と呼ばれる、国語研 NPCMJ コーパスの元となった句構造文法のツリーバンクを用いた。けやきツリーバンクは、基本的に句構造文法に基づくものの、(主語・目的語などの文法役割など)構造に直接反映されない文法情報も細かくタグ付けしてある点に特徴があり、言語学的に妥当な範疇文法のツリーバンクを構築するための基盤として特にふさわしい。本研究では、以下「研究の方法」で説明する方法により、このけやきツリーバンクを半自動で範疇文法のツリーバンクに変換し、変換したツリーバンクを web 上でオープン・ソースで公開した。

３．研究の方法

(1) ツリーバンク構築の手法

句構造文法のツリーバンクを範疇文法のツリーバンクに変換する試みとしては、Hockenmaier & Steedman (2007)、植松ら(2013)、Moot (2015)などによる先行研究があり、基本的手法はすでに確立している。本研究においても、変換手順の大枠に関しては、これらの先行研究を踏襲した。

本研究においては、特に、文末の複雑述語 (complex predicate) の扱いが言語学的に妥当なものとなるように注意を払った。複雑述語は、日本語だけでなく、ドイツ語、韓国語、ロマンス諸語など幅広い言語で見られる言語的特徴であり、カテゴリ文法を採用したツリーバンクにおいて言語学的に妥当な分析を施したアノテーションを行うことの意義が特に大きい。具体的なツリーバンク構築の手順としては、上述の先行研究を踏襲するパイプラインに、部分木の構造を修正するスクリプトを接ぎ木する形で変換プログラムを構成した。接ぎ木したスクリプトが本研究の要となる部分であり、この設計とエラー・チェックが、ツリーバンク構築の全作業のうちでも特に労力と時間を要した点である。これに関しては、研究期間の前半に研究代表者が変換スクリプトのプロトタイプを作成して、大まかな設計と遂行可能性の検証を行い、その後作業補助者のアノテータ二名が作業を引きつぎ、変換工程をより精巧にするという二段構えの工程で作業を行った。

(2) 研究遂行の手順

研究計画当初に決めた年次計画は以下の通り: ①1 年目:　ツリーバンクの荒削りな原型を作り、問題点を洗い出す。②2 年目: 問題点の集中的な修正。③3 年目: 仕上げ。パーザでの検証、成果を論文にまとめる。概ねこの年次計画に沿った形で研究を進めた。

(1)で述べたように、日本語の言語現象の扱いが適切なものとなるようなスクリプトの作成が本研究の眼目である。この段階においては、言語学の専門知識を持った作業者による丁寧なエラーチェック・修正作業が欠かせないものであった。この目的のため、アノテータとして言語学専攻の大学院博士(後期)課程の学生二名に謝金業務を依頼した。研究の実質的な内容と作業

方針に関して、窪田、峯島、アノテータの４名で定期的に打ち合わせをすることで研究を進めた。研究の進捗や問題への対処に関しては窪田が最終的な責任を持った。窪田、峯島、アノテータの大まかな役割分担は以下の通りとした。①窪田（代表者）：プロジェクトの管理、アノテータの監督、作業方針の策定、自動変換スクリプトのプロトタイプ作成。②峯島（分担者）：パーザでの検証、意味解析用のスクリプトの作成、アノテータへの助言・指導。③アノテータ：自動変換スクリプトの修正、アノテーションのエラーチェック。

４．研究成果

2021 年 3 月に、当初の計画通り ABC ツリーバンクを CC BY 4.0 で github 上に公開した。また、研究発表欄に記載した国内学会 2 件、国際学会 1 件で成果を発表し、予稿集論文を発表した。ツリーバンク構築に際して、句構造文法ツリーバンクを範疇文法ツリーバンクに自動変換するスクリプトを開発したため、これも関連資源として公開予定である。また、自然言語処理の意味解析に利用可能な言語資源であることを確認するため、研究期間の最後にパーザの学習と意味解析システムへの応用のパイロット実験を行った。パーザのモデルも github 上に公開した。

本研究で開発したツリーバンクは、日本語を対象とした深い意味解析を伴う自然言語処理研究のための研究資源として幅広く利用されることが期待される。特に、研究代表者らは 2021 年度から新たに「統語変換」の概念を組み込んだパーザの開発に取り組む科研課題を開始しており、ABC ツリーバンクで訓練したパーザを、より高度な言語分析を行うパーザのコンポーネントの一つとして利用することを計画している。

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　4件）

１．著者名 Yusuke Kubota, Koji Mineshima, Noritsugu Hayashi, Shinya Okano	４．巻 －
２．論文標題 Development of a General-Purpose Categorial Grammar Treebank	５．発行年 2020年
３．雑誌名 LREC 2020	６．最初と最後の頁 5195-5201
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 窪田悠介，峯島宏次，林則序，岡野伸哉	４．巻 －
２．論文標題 ABCツリーバンク：学際的な言語研究のための基盤資源	５．発行年 2021年
３．雑誌名 言語処理学会第27回年次大会予稿集	６．最初と最後の頁 1529-1534
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yusuke Kubota	４．巻 －
２．論文標題 A non-ellipsis-containing analysis of 'ellipsis-containing antecedents'	５．発行年 2021年
３．雑誌名 言語研究の楽しさと楽しみ: 伊藤たかね先生退職記念論文集	６．最初と最後の頁 264-274
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 窪田悠介、峯島宏次、林則序、岡野伸哉	４．巻 -
２．論文標題 汎用的な範疇文法ツリーバンクの構築	５．発行年 2019年
３．雑誌名 言語処理学会第25回年次大会予稿集	６．最初と最後の頁 143-146
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 窪田悠介、峯島宏次	４．巻 -
２．論文標題 前提投射の実例の ツリーバンクによる検索	５．発行年 2018年
３．雑誌名 日本言語学会第157回大会予稿集	６．最初と最後の頁 282-287
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計5件（うち招待講演　1件／うち国際学会　0件）

１．発表者名 窪田悠介
２．発表標題 カテゴリ文法・意味計算・文処理
３．学会等名 日本言語学会第158回大会ワークショップ「計算心理言語学--概要と展望--」
４．発表年 2019年

１．発表者名 窪田悠介
２．発表標題 理論言語学に未来はあるか?
３．学会等名 日本言語学会第159回大会シンポジウム「AIによって大きく揺さぶられる言語理論 --意味論の観点から--」（招待講演）
４．発表年 2019年

１．発表者名 窪田悠介，峯島宏次
２．発表標題 前提投射の統語コーパスでの検索
３．学会等名 関西言語学会第44回大会シンポジウム「高度文法情報付きコーパスとその日本語研究への応用」
４．発表年 2019年

１．発表者名 窪田悠介、峯島宏次、林則序、岡野伸哉
２．発表標題 汎用的な範疇文法ツリーバンクの構築
３．学会等名 言語処理学会第25回年次大会
４．発表年 2019年

１．発表者名 窪田悠介、峯島宏次
２．発表標題 前提投射の実例の ツリーバンクによる検索
３．学会等名 日本言語学会第157回大会
４．発表年 2018年

〔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
研究分担者	峯島　宏次 (Mineshima Koji) (80725739)	慶應大学・文学部・准教授 (32612)	

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