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　５ 年　 ６ 月 ２８ 日現在

機関番号：１２６１１
研究種目：基盤研究(B)（一般）
研究期間：2018～2021
課題番号：１８Ｈ０３２８４

研究課題名（和文）日本語CCG統語解析器lightblueの開発

研究課題名（英文）Development of Japanese CCG parser "lightblue"

研究代表者
戸次　大介（Bekki, Daisuke）

お茶の水女子大学・基幹研究院・教授

研究者番号：９０４３１７８３

交付決定額（研究期間全体）：（直接経費）　　13,200,000円

研究成果の概要（和文）：本研究は、理論言語学と深層ニューラルネットのハイブリッド手法によるCCG統語解析器の開発を目的とし、RNNGの文法理論をCFGからCCGに換装したシステムであるRNN-CCGの設計と実装に成功した。また、依存型意味論(DTS)による自動定理証明器の開発も行い、R2年度にはHaskellによる証明探索アルゴリズムを実装することに成功した。並行して、大規模言語モデルによる自然言語推論の性能を評価する研究や、比較構文、Weak Crossover、Proviso問題等、依存型意味論を用いた理論言語学の経験的研究においても成果を挙げた。

研究成果の学術的意義や社会的意義
本研究では、理論言語学と機械学習を融合させた深い意味解析を行う手法や、自然言語理解の自動化を可能にするCCGパーザと依存型意味論(DTS)の研究など、自然言語処理における重要な課題に取り組んできました。これらの研究成果は、査読付き国際学会や国内学会の発表を通じて学術的に評価され、また企業向けのセミナーやメディア掲載などを通じて社会に還元されました。今後、自然言語処理のさらなる進展に向けて、理論言語学と深層学習のハイブリッドアプローチを進めることが求められます。

研究成果の概要（英文）：The aim of this research was to develop a CCG syntactic parser based on a hybrid method of theoretical linguistics and deep neural networks. We successfully formulated and implemented RNN-CCG, a system that converts the grammatical theory of RNNG from CFG to CCG. We also studied automatic theorem prover with Dependent-Type Semantics (DTS) and successfully implemented a proof search algorithm by Haskell. In parallel, we also conducted research to evaluate the performance of natural language inference tasks with large-scale language models and showed their limitations, and empirical research in theoretical linguistics using dependent tye semantics, such as comparative constructions, Weak Crossover and the Proviso problem.

研究分野：計算言語学

キーワード：計算言語学　組合せ範疇文法　統語解析　深層ニューラルネットワーク

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

１．研究開始当初の背景

深層学習の新たな技術として、7 月に Allen Institute によって Elmo が、また 10 月に Google AI によって BERT が発表されたことにより、ニューラル自然言語処理は新たな段階に突入した。本研究は、形式文法理論とニューラルネットの融合を目指しているが、一部には BERT の登場に至って、形式文法理論の役割はニューラルネットによって完全に取って替わられた、という見解も散見された。したがって、それらの研究の限界点を見極める研究が本研究を推進する上で不可欠となった。

２．研究の目的

本研究では、日本語 CCG 形態素解析器＋統語解析器＋推論システムである lightblue の改良を加速的に推進し、日本語意味論データセット JSeM を対象とした含意関係認識を実現することを目指す。これまでの研究で、BERT を利用したニューラルネットで捉えうる統語的・意味的情報には一定の限界があることが明らかとなっており、本研究が目指す形式文法とニューラルネットの融合の重要性はますます高まるものと考えられる。

３．研究の方法

lightblue の設計は最先端の理論言語学の成果に基づいて設計されており、統語理論として組合せ範疇文法(CCG)を採用した頑健で高速な解析器であると同時に、意味理論として依存型意味論(DTS)を採用し、自然演繹に基づく証明探索アルゴリズムによって統語解析結果間の推論が計算可能である。また、形式文法理論と深層ニューラルネット(DNN)が融合した設計により、現在 DNN 単独では難しい「深い意味解析」へ到達することを目指す。また、lightblue は最新言語学理論のシミュレータとも見なせるため、本研究は理論言語学の検証可能性を引き上げる学際的研究プログラムとしての意義も併せ持つ。

４．研究成果

平成 30 年（令和 1 年）度

Elmo、BERT 等の言語モデルとの比較研究は当初の計画範囲を超える広がりをみせ、第一には、ニューラルネットが人間の推論の体系性を獲得しうるか、という一般的な問いに答えんとする研究に発展した。その成果は、[Yanaka+2019a] (*SEM2019 ワークショップ[査読付き国際学会])、[Yanaka+2019b] (2nd BlackBoxNLP (ACL2019 併設) ワークショップ[査読付き国際学会])、[谷中+2019] (第 33 回人工知能学会全国大会)、[谷中+2020] (言語処理学会第 26 回年次大会) において発表することができた（ともにオランダ・フローニンゲン大学との国際共同研究)。

また、第二には、本研究が与する理論言語学と機械学習のハイブリッドによる「深い意味解析」と、BERT のような深層学習のみによるアプローチの比較を行った。両アプローチの差が顕著となる言語現象として比較構文の研究を進めた。この研究では、CCG 統語解析器そのものを改良する代わりに、既存の CCG 統語解析器のナイーブな出力を Tsurgeon (Stanford NLP ツールに含まれる木構造変換プログラム) によって統語論的に妥当な構造に変換し、高度な意味解析に接続するという手法を採用した。この研究成果は、[Haruta+2020] (ACL student workshop (ACL-SRW2020)[査読付き国際学会]トップカンファレンス)、[Haruta+2019] (PACLIC33[査読付き国際学会])、[春田+2020] (人工知能学会第 34 回年次大会) において発表した。

本研究に注目して頂く機会も増え、2019 年度には国際学会で 2 回、国内学会で 1 回の招待講演に加えて、企業向けのセミナーで 2 回の一般向け講演を行い、研究成果の社会還元に務めた。

令和 2 年度

自然言語推論システム ccg2lambda は、本研究の前駆体に相当する研究成果であり、組合せ範疇文法(CCG)による統語解析、高階論理による意味合成、定理自動証明器を組み合わせた自然言語推論システムである。ニューラル自然言語処理の研究が進展をみせる中で、自然言語理解における限界も明らかになりつつあり、ccg2lambda のような理論言語学と深層学習のハイブリッドアプローチの重要性はますます増しつつある。

令和 2 年度は、(1) 形式統語論および形式意味論の最新の知見に基づく構造的言語処理として、ccg2lambda による比較構文の論理推論の研究、および (2)ccg2lambda に替わる意味の理論である依存型意味論(DTS)による自動定理証明アルゴリズムの開発に取り組んだ。

(1) ccg2lambda の利点として、標準的な形式意味論の分析に基づいた深い意味解析が可能であることが挙げられるが、一方で形式意味論の分析は必ずしも統一されておらず、異なる言語現象に対して異なる枠組みに基づく分析がなされていることがある。たとえば副詞の意味は event 意味論に基づいて分析され、比較構文は degree 意味論に基づいて分析されるが、本研究では、

副詞の比較構文という両者が相互作用する構文を取り上げ、event 意味論と degree 意味論の統一理論を設計し、ccg2lambda 上で実装するとともに、言語的に困難な問題を含む様々な NLI データセットを用いて評価した。その結果、従来の論理ベースのシステム、および深層学習ベースのシステムと比較して、高い精度を達成した。

(2) 依存型意味論による自動定理証明：自然言語の意味の理論として有力な依存型意味論(DTS)は、ΠΣ断片については、すでに型推論アルゴリズム[Bekki and Sato 2015]が存在するが、R2 年度の研究では Haskell による証明探索アルゴリズムを実装することに成功した。

コロナ禍によって予定していた学会参加による情報収集、学会発表による成果発表が実現できない時期が続いたが、一方で、3 件の査読付き国際学会論文発表、2 件の国内学会発表、2 件の国際学会招待講演、1 件のメディア記事（日経サイエンス）があり、おおむね順調な成果を挙げたと考えられる。

令和 3 年度

いよいよ理論言語学と深層学習のハイブリッドアプローチとして、1) 日本語 CCG パーザ lightblue に、ニューラル言語モデルによる形態素解析器を組み合わせる研究を行った。深層学習フレームワーク hasktorch（libtorch の Haskell バインディング）を用いて、既存の日本語形態素解析器の蒸留によって、少ない教師データから軽量なニューラル形態素解析器が得られることを示した[田上・戸次 2021]（6 月の人工知能学会において発表）。また、2) lightblue の意味計算部門で採用している自然言語の意味論のフレームワークである依存型意味論(DTS)を用いた理論言語学の研究として、Weak Crossover の研究[Bekki 2021]と proviso problem の研究[Yana+2021]を行った。それぞれ LENLS18 国際学会、LACL 国際学会において論文が採択され、発表を行った。その他、3) 日本語の実テキストに現れる数量詞の推論についての研究[小谷野・谷中・峯島・戸次 2021]、および 4) 意味論テストセットによる文法開発プラットフォームの確立を目指して、lightblue パーザ出力の可視化の研究[石嶋・戸次 2021]を進めた（ともに 6 月の人工知能学会において発表）。

当初の研究計画で述べたように、このプロジェクトの成果である CCG パーザ lightblue＋DTS prover という組合せは、形式統語論と形式意味論の検証過程の自動化と見做すこともできる。そのように統合され自動化された言語理論が、既存の理論言語学と比較して、はたまた自然言語処理における意味解析と比較して、何を意味するのか。そもそも人間の言語機能を科学的に研究するとはどのような行為であるのか。2021 年度言語学フェスでの発表[戸次 2021]では、そのような問題意識と、このプロジェクトを通して得られつつある回答を素描した。

令和 3 年度については、当初の研究計画では、前年度から繰り越した課題である CCG パーザとニューラルネットの融合研究を進め、8 月に ESSLLI サマースクールにおける研究発表と情報収集を行い、11 月には LENLS18 国際ワークショップにおける併設シンポジウムの形式で、4 年間の成果の取りまとめを行う予定であった。しかしながら新型コロナウィルス感染拡大に伴い、2019 年度から延期を重ねていた ESSLLI サマースクールも、LENS18 もともにオンライン開催となり、想定していたような情報収集や宣伝を含めた成果発表の機会には恵まれなかった。そのような状況下ではあったが、2 件の査読付き国際学会論文発表、3 件の国内学会発表、2 件の招待講演があり、おおむね順調な成果を挙げたと考えられる。

令和 4 年度（繰越による）

当初計画の目標であった CCG パーザとニューラルネットの融合研究について、前述の計画変更を経て、最終的にはプロジェクトの集大成といえる RNN-CCG の枠組みの実装に成功し、成果を論文にまとめることができた。RNN-CCG とは、Recurrent Neural Network Grammar (RNNG)が Context-Free Grammar (CFG)に基づいているのに対し、CFG を組み合わせ範疇文法(CCG)に変更した枠組みである。この枠組みは、当初この研究が目指していたニューラル CCG 統語解析を昇華したものであり、成果は[田上・戸次 2023]として令和 5 年 3 月に研究発表を行うとともに、現在国際学会に論文を投稿中である。

また、CCG パーザと両輪をなす依存型意味論(DTS)の研究成果としては、Weak Crossover の研究[Bekki2021]、Proviso 問題の研究[Yana+2021]、Comparative 構文の研究[Haruta+2023]を行い、それぞれ査読付き国際学会予稿集、査読付き国際ジャーナルに論文が掲載された。

また新型コロナウィルス感染症による影響で延期となった ESSLLI2021 であったが、2022 年度に ESSLLI2022 が現地開催されたため、予定していた交流・情報収集を行うことができた。また、久しぶりの現地開催となった令和 4 年 11 月の LENLS19 において、当初予定していた招待講演者の招聘を行うことができた。その他、国際学会における招待講演が 3 本、国内学会・研究会における招待講演が 2 本、国内学会における発表が 3 本あり、研究成果の普及に努めた。

５．主な発表論文等

〔雑誌論文〕　計11件（うち査読付論文　11件／うち国際共著　3件／うちオープンアクセス　3件）

１．著者名 Yukiko Yana, Koji Mineshima, Daisuke Bekki	４．巻 X
２．論文標題 The proviso problem from a proof-theoretic perspective	５．発行年 2021年
３．雑誌名 Logical Aspects of Computational Linguistics (LACL) 2021	６．最初と最後の頁 159-176
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 Daisuke Bekki	４．巻 X
２．論文標題 A Proof-theoretic Analysis of Weak Crossover	５．発行年 2021年
３．雑誌名 the 18th International Workshop on Logic and Engineering of Natural Language Semantics (LENLS18)	６．最初と最後の頁 75-88
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 Haruta,Izumi; Mineshima,Koji; Bekki,Daisuke;	４．巻 X
２．論文標題 Combining Event Semantics and Degree Semantics for Natural Language Inference	５．発行年 2020年
３．雑誌名 Proceedings of the COLING2020 (short paper), Barcelona, Spain (Online)	６．最初と最後の頁 1758-1764
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/2020.coling-main.156	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 Haruta,Izumi; Mineshima,Koji; Bekki,Daisuke;	４．巻 X
２．論文標題 Logical Inferences with Comparatives and Generalized Quantifiers	５．発行年 2020年
３．雑誌名 Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics Student Research Workshp (ACL2020-SRW), Seattle, USA.	６．最初と最後の頁 263-270
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/2020.acl-srw.35	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 Yana Yukiko、Mineshima Koji、Bekki Daisuke	4．巻 28
2．論文標題 Variable Handling and Compositionality: Comparing DRT and DTS	5．発行年 2019年
3．雑誌名 Journal of Logic, Language and Information	6．最初と最後の頁 261～285
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10849-019-09294-3	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 Haruta, Izumi; Mineshima, Koji; Bekki, Daisuke;	4．巻 －
2．論文標題 Logical Inferences with Comparatives and Generalized Quantifiers	5．発行年 2020年
3．雑誌名 Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics Student Research Workshp (ACL2020-SRW)	6．最初と最後の頁 －
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 Haruta, Izumi; Mineshima, Koji; Bekki, Daisuke;	4．巻 －
2．論文標題 A CCG-based Compositional Semantics and Inference System for Comparatives	5．発行年 2019年
3．雑誌名 Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC33)	6．最初と最後の頁 67-76
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 Yanaka Hitomi、Mineshima Koji、Bekki Daisuke、Inui Kentaro、Sekine Satoshi、Abzianidze Lasha、Bos Johan	4．巻 －
2．論文標題 Can Neural Networks Understand Monotonicity Reasoning?	5．発行年 2019年
3．雑誌名 Proceedings of the Second BlackboxNLP workshop on Analyzing and Interpreting Neural Networks for NLP	6．最初と最後の頁 31-40
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/W19-4804	査読の有無 有
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 該当する

1．著者名 Yanaka Hitomi、Mineshima Koji、Bekki Daisuke、Inui Kentaro、Sekine Satoshi、Abzianidze Lasha、Bos Johan	4．巻 －
2．論文標題 HELP: A Dataset for Identifying Shortcomings of Neural Models in Monotonicity Reasoning	5．発行年 2019年
3．雑誌名 Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)	6．最初と最後の頁 250-255
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/S19-1027	査読の有無 有
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 該当する

1．著者名 Kubota Yusuke、Mineshima Koji、Levine Robert、Bekki Daisuke	4．巻 －
2．論文標題 Underspecification and interpretive parallelism in Dependent Type Semantics	5．発行年 2019年
3．雑誌名 Proceedings of the IWCS 2019 Workshop on Computing Semantics with Types, Frames and Related Structures (CSTFRS)	6．最初と最後の頁 1-9
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/W19-1001	査読の有無 有
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 該当する

1．著者名 Watanabe Kazuki、Mineshima Koji、Bekki Daisuke	4．巻 －
2．論文標題 Questions in Dependent Type Semantics	5．発行年 2019年
3．雑誌名 Proceedings of the Sixth Workshop on Natural Language and Computer Science (NLCS'19)	6．最初と最後の頁 23-33
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/W19-1103	査読の有無 有
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 －

〔学会発表〕　計21件（うち招待講演　8件／うち国際学会　6件）

1．発表者名 Daisuke Bekki
2．発表標題 A Proof-theoretic Analysis of Weak Crossover
3．学会等名 the 2nd workshop on Language Faculty Science (2022/2/12)（招待講演）（国際学会）
4．発表年 2022年

１．発表者名
戸次大介

２．発表標題
言語理論の証明論的転回

３．学会等名
言語学フェス2022（2022/1/29）

４．発表年
2022年

１．発表者名
田上青空，戸次大介

２．発表標題
日本語形態素解析器の知識蒸留

３．学会等名
第34回人工知能学会全国大会

４．発表年
2021年

１．発表者名
石嶋美咲，戸次大介

２．発表標題
Yesodによる日本語CCGパーザ開発環境の構築

３．学会等名
第34回人工知能学会全国大会

４．発表年
2021年

１．発表者名
小谷野華那，鈴木莉子，春田和泉，谷中瞳，戸次大介

２．発表標題
実テキストにおける数量表現の含意関係認識に向けて

３．学会等名
第34回人工知能学会全国大会

４．発表年
2021年

１．発表者名
戸次大介

２．発表標題
理論言語学と深層学習のハイブリッドアプローチによる自然言語推論

３．学会等名
半導体エネルギー研究所（2021/5/19）（招待講演）

４．発表年
2021年

１．発表者名
大洞日音，戸次大介

２．発表標題
DTSの部分体系のための定理自動証明器の実装に向けて

３．学会等名
言語処理学会第27回年次大会，北九州国際会議場／オンライン，2021/3/15-19.

４．発表年
2021年

１．発表者名
Bekki,Daisuke

２．発表標題
Why parsing is a part of Language Faculty Science

３．学会等名
The 2020 Zoom Workshop on Language Faculty Science: Linguistic Intuitions and Replication, 2020/12/20（招待講演）（国際学会）

４．発表年
2020年

１．発表者名
Bekki,Daisuke

２．発表標題
A hybrid approach toward Natural Language Understanding

３．学会等名
Centre for Linguistic Theory and Studies in Probability (CLASP), 2020/12/09（招待講演）（国際学会）

４．発表年
2020年

１．発表者名
Daido,Hinari; Bekki,Daisuke;

２．発表標題
Development of an automated theorem prover for the fragment of DTS

３．学会等名
the 17th International Workshop on Logic and Engineering of Natural Language Semantics (LENLS17).（国際学会）

４．発表年
2020年

１．発表者名
春田和泉，峯島宏次，戸次大介

２．発表標題
CCGと自動定理証明による比較表現の計算意味論

３．学会等名
人工知能学会第34回年次大会，オンライン開催，2020/6/9-12.

４．発表年
2020年

１．発表者名
谷中瞳，峯島宏次，戸次大介，乾健太郎

２．発表標題
ニューラルネットは自然言語推論の体系性を学習するか

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
伊藤美賀，佐藤七海，田上青空，谷中瞳，峯島宏次，戸次大介

２．発表標題
汎用言語モデルBERTを用いた多言語テキストにおける意味現象タグ予測

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
秋山雛乃，石嶋美咲，石田真捺，高野紗輝，鈴木莉子，谷中瞳，峯島宏次，戸次大介

２．発表標題
CCGとCoqを用いた日本語マルチモーダル推論システムの構築

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
飯野早貴，石田真捺，小谷野華那，松本留奈，鈴木莉子，谷中瞳，峯島宏次，戸次大介

２．発表標題
マルチモーダル推論評価のための日本語データセットの試案

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
谷中瞳，戸次大介，峯島宏次，関根聡，乾健太郎

２．発表標題
クラウドソーシングによる単調推論データセットの構築

３．学会等名
第33回人工知能学会全国大会

４．発表年
2019年

１．発表者名
鈴木莉子，吉川将司，谷中瞳，峯島宏次，戸次大介

２．発表標題
テキスト情報と画像情報を組み合わせた論理推論システムの構築

３．学会等名
第33回人工知能学会全国大会

４．発表年
2019年

１．発表者名
Bekki, Daisuke

２．発表標題
Dependent Types and Theory of Meaning

３．学会等名
Nanzan Workshop on the Foundational Issues in Linguistics and Philosophy of Language（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
戸次大介

２．発表標題
理論言語学に基づく自然言語理解の最前線

３．学会等名
日本英語学会第37回大会（招待講演）

４．発表年
2019年

１．発表者名
Bekki, Daisuke; Yanaka, Hitomi

２．発表標題
Hybrid natural language understanding: neural network, logic and beyond

３．学会等名
Konferenz zur Verarbeitung naturlicher Sprache (KONVENS2019)（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
戸次大介

２．発表標題
自然言語理解技術の動向

３．学会等名
NEDO TSC Foresightセミナー（第1回）（招待講演）

４．発表年
2019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

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2件

国際研究集会 Logic and Engeneering of Natural Language Semantics (LENLS18)	開催年 2021年～2021年
国際研究集会 Logic and Engeneering of Natural Language Semantics (LENLS17)	開催年 2020年～2020年

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			
オランダ	フローニンゲン大学			