言語学的に妥当な日本語 CCG ツリーバンクの構築と評価

富田朝 [1]　谷中瞳 [2]　戸次大介 [1]
[1] お茶の水女子大学　[2] 東京大学
{tomita.asa,bekki}@is.ocha.ac.jp
hyanaka@is.s.u-tokyo.ac.jp

概要

日本語 CCG パーザが日本語を正しく分析するためには、パーザの学習・評価に用いられる日本語 CCG ツリーバンクの言語学的妥当性を向上させる必要がある。しかし、既存の日本語 CCG ツリーバンクである日本語 CCGbank には誤った分析が含まれていることが指摘されており、日本語 CCG ツリーバンクの新たな構築アルゴリズムが提案されている。本研究では、CCG ツリーバンクを構築するアルゴリズムを実装し、特に複合動詞を含む文に対して正しい統語構造を出力できるようにアルゴリズムを改良した。さらに、13,653 文からなる日本語 CCG ツリーバンクである lightblue CCGbank を構築し、ツリーバンクの統語構造と意味表示に対して、人手で評価を行った。

1　はじめに

ツリーバンクは、大規模なテキストに統語構造が付与されたデータセットである。特に形式統語論に基づいたツリーバンクは、パーザの学習・評価用データセットとしても利用され、パーザ技術の発展に貢献している。形式統語論の理論に基づいて構築された英語のツリーバンクとして、文脈自由文法に基づいた Penn Treebank [1] や、組合せ範疇文法 (Combinatory Categorial Grammar, CCG [2, 3]) に基づいた CCGbank [4] などが挙げられる。また、日本語のツリーバンクとしては、AB 文法に合成規則を加えた理論（ABC 文法）に基づいた ABC ツリーバンク [5] や、句構造文法に基づいた係り受け構造のツリーバンクから自動変換することによって構築された日本語 CCGbank [6] などがある。とくに、日本語 CCGbank の構築は、日本語 CCG パーザの研究にも影響を与えており、Jigg [7] や depccg [8] などの日本語 CCG パーザが開発されるに至った。さらに、これらの CCG パーザは ccg2lambda [9] をはじめとした、意味解析・推論システムに活用されている。

近年では深層学習を用いた手法の発展によって高精度な CCG パーザが実現しつつあるが、CCG パーザの精度が高いということが、そのパーザが言語学的に妥当な統語構造を出力していることを意味するわけではないことに注意したい。ここで、言語学的に妥当な統語構造とは、日本語文法の形式理論 [10] に基づき、構成的に正しい意味表示を導出できるような統語構造を指す。CCG パーザの評価は、CCG ツリーバンクの評価用データをどれくらい再現できるか、という観点から行われる。しかし、正解とされる CCG ツリーバンクが言語学的な誤りを含んでいる場合、CCG ツリーバンクの統語構造を再現できるほど評価値としてのパーザの精度は高くなるが、そのパーザが言語学的に妥当な統語構造を出力できているとはいえない。実際に、既存の日本語 CCG パーザの学習・評価データとして使われている日本語 CCGbank は、受身・使役の分析に誤りが含まれていることが指摘されている [11]。したがって、言語学的に妥当な CCG パーザの実現には言語学的に妥当な CCG ツリーバンクの構築が必要となる。

富田ら [12] は、CCG に基づく統語・意味解析器である lightblue [13] と ABC ツリーバンクを組み合わせることで言語学的に妥当な日本語 CCG ツリーバンクを構築するアルゴリズムを提案している。しかし、この研究で提案されたアルゴリズムにはいくつかの課題が残されている。そこで本研究では、富田ら [12] のアルゴリズムの課題のうち、複合動詞の解析に焦点をあてて、アルゴリズムの改良を行う。そして、改良したアルゴリズムに基づいてツリーバンクを再構築する。再構築によって得られたツリーバンクは、日本語 CCG 統語・意味解析器 lightblue を用いて得られたことから、lightblue CCGbank と名付ける。さらに、lightblue CCGbank の統語構造と意味表示について評価を行い、言語学的な妥当性の考察とエラー分析を行う。

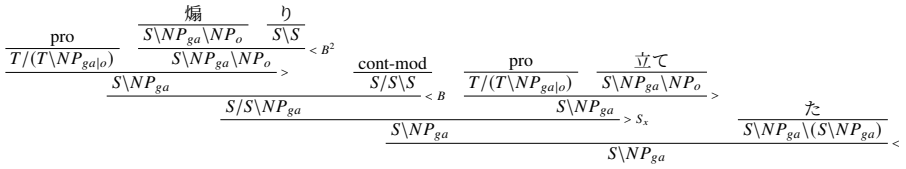

図 1　複合動詞「煽り立て」の統語構造

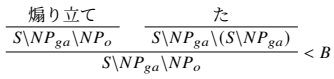

図 2　アルゴリズム改良後の複合動詞「煽り立て」の統語構造

2　背景

2.1　lightblue の基本仕様

lightblue [13] は CCG による日本語の文法理論 [10] に基づいて開発された統語・意味解析器である。lightblue は格フレームを付与した Juman[14] の辞書約 8 万語に、機能語約 800 語を加えた辞書と、CCG の組合せ規則に基づいて統語解析を行う。受け取った入力文の部分文字列をなす全ての語彙項目を辞書から取得し、それらを用いて Left-corner chart parsing [15] を行う。その後、CCG の統語構造と依存型意味論 (Dependent Type Semantics, DTS) [16] に基づく意味表示を解析結果として出力する。

lightblue で使用されている用言の語彙項目の格フレームは、コーパスから自動獲得されたものであるため、誤りが含まれている。したがって、lightblue の格フレームの誤りをどのように減らすかが課題として残されている。

2.2　富田らによる提案手法

富田ら [12] では、ABC ツリーバンク [5] と lightblue を用いて日本語 CCG ツリーバンクを構築する手法が提案された。ABC ツリーバンクは、句構造文法で記述されているけやきツリーバンク[1] を変換して構築されたツリーバンクであり、項構造をはじめとした統語情報が人手でアノテーションされている。自動変換のみによって構築されたツリーバンクと比較して、言語学的に妥当な統語情報が多く含まれていることが特徴として挙げられるが、品詞や活用形に関する統語情報が不十分であるという課題がある。

そこで富田ら [12] の手法では、lightblue の用言の語彙項目に含まれる格フレームの誤りを減らすために、ABC ツリーバンクの用言の項構造で lightblue の辞書の情報を上書きする。具体的にはまず、ABC ツリーバンクから、文とその文に含まれる用言の統語情報を抽出する。その後、抽出した文を lightblue に入力し、統語解析を行う。獲得した語彙項目の一部を、ABC ツリーバンクから抽出した用言の統語情報で上書きすることで語彙項目をアップデートする。最後に、新しい語彙項目を用いて lightblue で入力文をパーズすることで、正しい統語構造と意味表示からなる出力が得られる。

3　ツリーバンク構築アルゴリズム

3.1　アルゴリズムの改良

富田ら [12] の手法では、(1) のような複合動詞を含む文において、図 1 のように、2 つの動詞が分割して解析されてしまうという課題がある。

(1)　急進派の発言は、国民の不安を煽り立てた[2]

複合動詞には、「煽り立てる」のようにまとめて一語とされるものと、「食べ始める」のように動詞と動詞性接尾語（補助動詞）の二語が合成されているものがある。機械的にこれらを区別することは難しいため、本研究では、全ての複合動詞について、一語となるようにアルゴリズムを改良する。

ABC ツリーバンクでは一語の複合動詞として扱われているが、lightblue の辞書には登録されていない複合動詞も存在する。このとき、アルゴリズム内で lightblue の語彙項目と ABC ツリーバンクの用言の表層形とのマッピングを適切に行うことができず、誤った解析が行われてしまう。そこで、複合動詞のリストをあらかじめ用意し、lightblue でパーズする際に以下の操作を行うことで、複合動詞を一語として解析できるように改良した。

1. ABC ツリーバンクから抽出した用言のリストの中に、複合動詞のリストに含まれている要素があるかを判定する
2. 複合動詞が存在した場合、lightblue の用言のテンプレートを用いて複合動詞の語彙項目を作成する
3. lightblue の chart parsing で用いる用言の語彙項目を、2 節で作成した語彙項目で上書きする

1）　https://github.com/ajb129/KeyakiTreebank

2）　ABCTreebankID:3_dict_vv-lexicon

4. 新しい情報で上書きされた語彙項目を用いて lightblue の chart parsing を行う

この操作を加えたアルゴリズムで (1) の「煽り立てた」をパーズした結果を図 2 に示す。「煽り立て」という複合動詞がガ格名詞とヲ格名詞を必須格とする一語の動詞として扱われるようになり、正しい統語構造を導出できているといえる。

しかし、(2) の文のように、正しい統語構造を得ることができないケースも存在する。(2) の統語構造は、付録の図 5 に示す。

(2)　彼はじっととまって、あたりを見廻した。[3)]

複合動詞「見廻る」を分割して解析すると、「彼がじっととまって、あたりを見」た後に「廻す」をいう動作をおこなった、という意味になる。「見廻す」の主語は彼であるにもかかわらず、この解析では「彼」のスコープは「見」までとなり、廻るの主語は空範疇となるため誤りである。この文を正しく解析できない原因については 4.2.2 節で詳しく述べる。

3.2　lightblue CCGbank の構築

ABC ツリーバンクのファイルはジャンル別に分類されており、それぞれのジャンルから最大 5 ファイルずつをランダムに抽出し、15,137 文を抽出した。lightblue の chart parsing は、$O(n^3)$ の計算量を要するため、文が長くなるとパーズに時間がかかる。そこで本論文では 50 文字以上の文は対象外とし、計 13,653 文のツリーバンクを構築した。なお、ランダムに抽出した 15,137 文のうち 50 文字以上の文は 1,482 文と全体の 10%に満たないため、前述の制約を設けても構築アルゴリズムのスケーラビリティは保証されると判断できる。抽出した文のジャンル別の分類は付録の表 2 に示す。

4　ツリーバンクの評価と考察

4.1　評価

ツリーバンクの言語学的な妥当性を自動で評価することは困難であるため、人手で評価する必要がある。しかし、CCG 統語構造や、DTS の意味表示を人手で評価することはコストの高い作業であり、今回得られた 13,653 文の全てについて評価を行うには膨大な時間を要する。そこで、lightblue CCGbank の各ジャンルから 4 文ずつをランダムに抽出して得られた 56 文に対して、人手で妥当性の評価を行った。

3)　ABCTreebankID: 1089_aozora_Harada-1960

表 1　ツリーバンクのエラーの分類

	分類	該当数
統語構造のエラー	未登録語、統語範疇	18
	複合動詞	4
	その他統語構造	30
意味表示のエラー	漢字のエラー	7

56 文のうち、統語構造と意味表示の両方が妥当であると判断できた木は 19 文で、全体の 33% であった。妥当でない文に対しては、その文に含まれるエラーを以下の 3 つに分類した。

1. 辞書に未登録の機能語があり、付与されている統語範疇が妥当でない
2. 複合動詞を分割して解析している
3. その他の統語構造のエラーが含まれている

また、意味表示の評価として、正しい漢字を割り当てられているか、という観点も加えて評価した。エラー分類の結果を表 1 に示す。

4.2　考察

4.2.1　受身、使役を含む文

日本語 CCGbank は、受身・使役の分析に誤りが含まれている。そこで本節では、lightblue CCGbank において受身・使役の分析がどのように改善されたのかについて議論する。lightblue CCGbank には、(3)、(4) のような受身・使役文が含まれている。

(3)　太郎は先生に絵をほめられた[4)]

(4)　私は猫に魚を食べさせた[5)]

これらの文に対応する CCG 統語構造を図 3 に示す。受身・使役の動詞性接尾語「れ」「せ」には、統語範疇 $S\backslash NP_{ga}\backslash NP_{ni}\backslash(S\backslash NP_{ga})$ が付与されており、[11] で指摘されていた日本語 CCGbank の課題を解決できていることがわかる。

4.2.2　エラー分析

本節では評価結果から特定された、統語構造に関する 3 種類のエラーについて、具体例を参照して議論する。

未登録の機能語に由来するエラー

(5)　国際関係の仕事に今ついてるのね。[6)]

(5) では、終助詞「ね」が「寝る」という動詞の語幹

4)　ABCTreebankID:413_textbook_kisonihongo
5)　ABCTreebankID: 693_textbook_purple_intermediate
6)　ABCTreebankID: 68_spoken_JF10

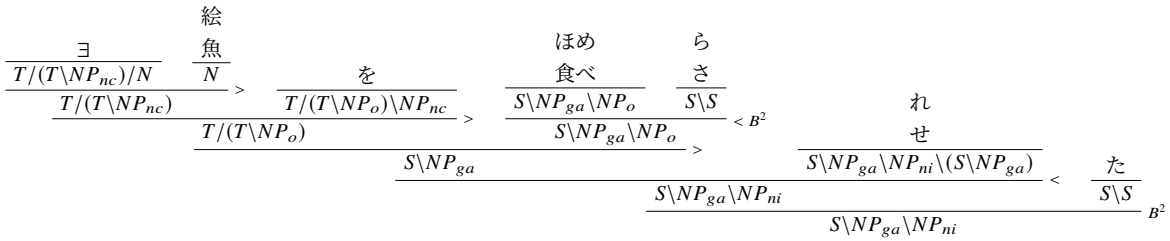

図 3　受身の文 (3) と使役の文 (4) の統語構造の一部

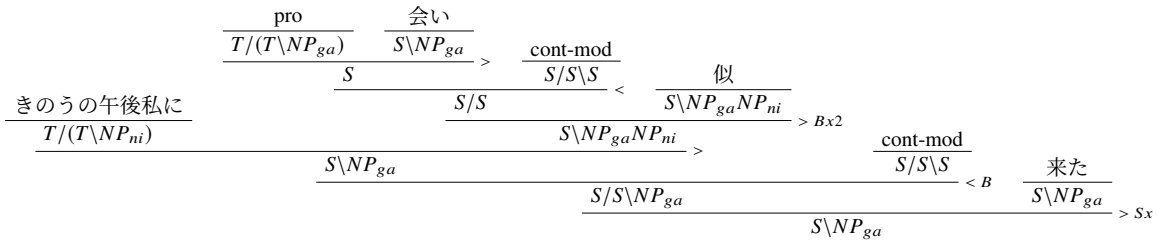

図 4　(6) の統語構造の一部

として解析された。これは lightblue の辞書に終助詞の「ね」が登録されていないことが原因と考えられる。このエラーを改善するためには、終助詞「ね」を lightblue の辞書に追加し、適切な制限を設ける必要がある。

複合動詞のエラー　複合動詞のエラーとして、(2) で紹介した「見廻し」に加えて、「消し去る」「言い聞かせ」「申し上げる」といった複合動詞の解析に誤りが含まれていた。これらは、図 1 のエラーのような動詞同士が連用節として結合されてしまうというエラーではなく、「見」と「廻し」が結合されるよりも前に、「見」の前にある「あたりを」というヲ格名詞と結合してしまうことで引き起こされているエラーであった。

エラーの原因として、改良したアルゴリズムによって「見廻し」という語彙項目を新しく追加することは実現できたものの、chart parsing の際に、新しく作成した語彙項目の優先順位が低くなってしまっていることなどが挙げられる。今後、さらなる分析を行い、すべての複合動詞に対して有効なアルゴリズムを構築していく必要がある。

統語構造のエラー　統語構造のエラーの例として、(6) が挙げられる。(6) の統語構造を図 4 に示す。

(6)　太郎は、きのうの午後私に会いに来た。[7)]

このエラーでは、「会いに来た」の格助詞「に」が、動詞の「似る」の語幹として解析されている。ひらがな一文字の助詞が動詞として扱われてしまうエラーは多く、一文字のひらがなに対して動詞を割り当てる際に制限を設ける必要がある。

5　おわりに

本研究では、言語学的に妥当かつ詳細な統語情報を有する日本語 CCG ツリーバンクの構築を目指して、複合動詞の扱いを中心にツリーバンクを構築するアルゴリズムの改良を行なった。改良したアルゴリズムを用いて 13,653 文の日本語 CCG ツリーバンクである lightblue CCGbank を構築し、得られた CCG 統語構造と DTS 意味表示について人手で評価を行った。評価の結果、既存の日本語 CCGBank の課題の一つである受身・使役に関して正しい分析に基づく CCG 統語構造を導出できている傾向が示唆された。一方で、今回改良したアルゴリズムはすべての複合動詞については有効でないことがわかった。

今後、ツリーバンク構築アルゴリズムのさらなる改良や、lightblue の辞書の拡張を進めることで、CCG ツリーバンクの改善を行う。また、構築した lightblue CCGbank は後日研究利用可能な形式で公開する予定である。

7)　787_textbook_kisonihongo

謝辞

本研究は JSPS 科研費 JP20K19868 の支援を受けたものである。

参考文献

[1] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. **Computational Linguistics**, Vol. 19, No. 2, pp. 313–330, 1993.

[2] Mark Steedman. **Surface Structure and Interpretation**. The MIT Press, Cambridge, 1996.

[3] Mark Steedman. **The Syntactic Process**. MIT Press, 2000.

[4] Julia Hockenmaier and Mark Steedman. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. **Computational Linguistics**, Vol. 33, No. 3, pp. 355–396, 2007.

[5] Yusuke Kubota, Koji Mineshima, Noritsugu Hayashi, and Shinya Okano. Development of a general-purpose categorial grammar treebank. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 5195–5201, Marseille, France, May 2020. European Language Resources Association.

[6] Sumire Uematsu, Takuya Matsuzaki, Hiroki Hanaoka, Yusuke Miyao, and Hideki Mima. Integrating multiple dependency corpora for inducing wide-coverage Japanese CCG resources. In **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1042–1051, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[7] Hiroshi Noji and Yusuke Miyao. Jigg: A framework for an easy natural language processing pipeline. In **Proceedings of ACL-2016 System Demonstrations**, pp. 103–108, Berlin, Germany, August 2016. Association for Computational Linguistics.

[8] Masashi Yoshikawa, Hiroshi Noji, and Yuji Matsumoto. A* CCG parsing with a supertag and dependency factored model. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 277–287, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[9] Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. ccg2lambda: A compositional semantics system. In **Proceedings of ACL-2016 System Demonstrations**, pp. 85–90, Berlin, Germany, August 2016. Association for Computational Linguistics.

[10] 戸次大介. 日本語文法の形式理論. くろしお出版, 2010.

[11] Daisuke Bekki and Hitomi Yanaka. Is Japanese CCGBank empirically correct? A case study of passive and causative constructions. In **Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)**, pp. 32–36, Washington, D.C., March 2023. Association for Computational Linguistics.

[12] 富田朝, 谷中瞳, 戸次大介. 言語学的に妥当な CCG ツリーバンク構築の試み. 人工知能学会全国大会論文集, Vol. JSAI2023, pp. 2E6GS605–2E6GS605, 2023.

[13] Daisuke Bekki and Ai Kawazoe. Implementing variable vectors in a CCG parser. In **Logical Aspects of Computational Linguistics. Celebrating 20 Years of LACL (1996–2016)**, pp. 52–67, Berlin, Heidelberg, 12 2016. Springer Berlin Heidelberg.

[14] Daisuke Kawahara and Sadao Kurohashi. Case frame compilation from the web using high-performance computing. In **Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)**, Genoa, Italy, May 2006. European Language Resources Association (ELRA).

[15] Klaas Sikkel. **Left-Corner chart parsing**, pp. 201–220. Springer Berlin Heidelberg, Berlin, Heidelberg, 1997.

[16] Daisuke Bekki and Koji Mineshima. Context-passing and Underspecification in Dependent Type Semantics. 2017.

付録

表 2　ツリーバンクのジャンルと文の数

ジャンル	ABC ツリーバンクの文の数	50 字以上の文	CCG ツリーバンクの文の数
aozora	1773	590	1183
bible	1652	220	1430
book_expert	50	4	41
dict_lexicon	2640	4	2636
diet_kaigiroku	486	112	374
fiction	921	44	877
law	337	128	209
misc	335	59	276
news	443	103	340
non-fiction	223	87	126
spoken	570	11	559
ted_talk	605	54	551
textbook	4880	10	4870
wikipedia	222	51	171
計	15137	1482	13653

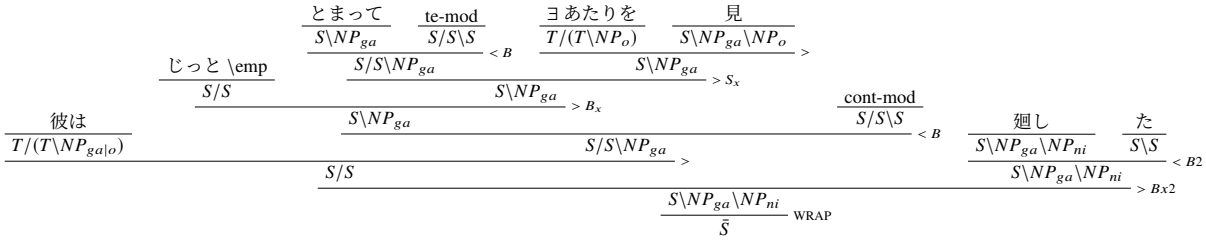

図 5　(2) の統語構造