CCGbank の言語学的妥当性の検証

富田 朝（指導教員：戸次大介）

1 はじめに

組合せ範疇文法 (Combinatory Categorial Grammar: CCG)[4] は語彙化文法の一種で，統語構造や意味に関わる情報のほとんどがあらかじめ辞書 (lexicon) に記述されている．さらに単純な規則を組み合わせることで文構造を記述することができる．CCG に基づいた統語構造が付与されたコーパスを CCG ツリーバンクといい，日本語の CCG ツリーバンクとして日本語 CCGbank[5] がある．

depccg[3] をはじめとした CCG パーザは，学習・評価データとして CCG ツリーバンクを利用することが多いため，CCG パーザの精度は CCG ツリーバンクの言語学的正しさに依存する．したがって CCG パーザの精度向上のため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．本稿では，言語学的に妥当な日本語 CCG ツリーバンクの構築に向けて，日本語 CCGbank の言語学的な問題点を解消した日本語 CCG ツリーバンクを生成する手法を提案する．

2 先行研究

現存の日本語 CCG ツリーバンクとして，日本語 CCGbank[5] がある．近年，より質の高いツリーバンクの構築を目指し ABC 文法に基づいた ABC ツリーバンク [7, 8] が構築された．

2.1 日本語 CCGbank

日本語 CCGbank は，係り受け構造で記述された京都大学テキストコーパス[1]，述語項構造や照応関係の情報が付与された NAIST コーパス[2], 助詞「と」を含む文の項と述語の関係の情報を付与した「と」コーパスから得られた統語情報を統合的に利用して構築された．しかし，日本語 CCGbank は受身・使役の構文に対して誤った分析がなされていることが指摘されている [2]．誤りが指摘されている受身文の例を図 1，日本語 CCG[6] に基づいた言語学的に妥当な分析を図 2 に示す．日本語 CCGbank では受身の接尾語「れ」の統語範疇が $S\backslash S$ と分析されているが，日本語 CCG[6] によると，これは $S\backslash NP_{ga}\backslash NP_{ni}\backslash(S\backslash NP_{ga}\backslash NP_{ni|o})$ と分析されるべきものである．

2.2 ABC ツリーバンク

ABC ツリーバンクは， 国立国語研究所の現代日本語統語意味コーパスの関連資源である，けやきツリーバンク[3]を ABC 文法による統語構造に変換することで構築された．ABC 文法とは，関数適用規則のみからなる AB 文法に関数合成規則 (function composition) を加えた文法で，範疇文法の中間言語的枠組みとされている．関数合成規則は CCG においても基本的な規則であるため ABC ツリーバンクは容易に CCG ツリーバンクへ変換することができるという特徴を持つ．

ABC ツリーバンクでは，日本語 CCGbank の受身・使役の構文の誤りをはじめとした，項構造などの課題が多く改善された．しかし，活用種別など日本語 CCG の標準的な統語素性に関する記述がなく，統語情報としては不十分であるという課題を残す．

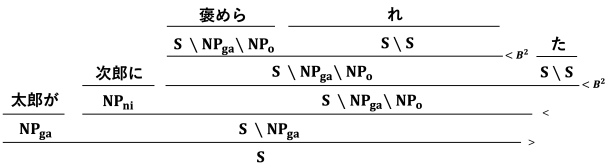

図 1: 日本語 CCGbank の CCG 木

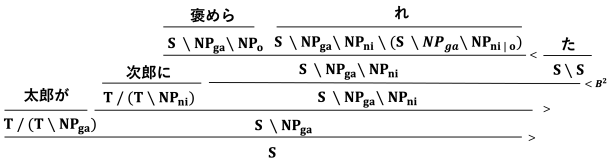

図 2: Bekki (2010) に基づいた CCG 木

2.3 lightblue

CCG 統語解析器 lightblue[1][4]は日本語 CCG における詳細な統語素性についての議論 [6] に基づいて構築された．lightblue は日本語 CCG[6] の形式で記述された統語構造を出力するが，項構造に関する誤りが多いという課題がある．これは lightblue の辞書が用いる格フレームがコーパスから自動獲得されたものであり，誤りを含むためである．その結果，lightblue は本来自然ではない用言の選択がなされていたり，選択された用言の格フレームに間違いが含まれていたりと，出力の統語構造の質に改善の余地がある．

3 提案手法

本研究では，項構造が人手によって正確に記述されている ABC ツリーバンクの特徴と，出力に詳細な統語素性を与えることができる lightblue の特徴を合わせることで，言語学的に妥当で詳細な情報を持ったツリーバンクを生成することを目標とする．まず lightblue の項構造の妥当性を向上させるために，ABC ツリーバンクから得られた用言の語彙項目を用いて lightblue の語彙項目をフィルターする．その後，語彙項目を修正した lightblue を用いて ABC ツリーバンクの文をパーズし，その結果をツリーバンクとして出力する．提案手法の流れを図 3 に示す．実装は Haskell 言語で行う．

3.1 ABC ツリーバンクのパーズ

まず ABC ツリーバンクのパーザを作成し，ABC ツリーバンクを木構造データとして扱えるようにする．その後，木構造データから統語範疇が用言である語のみを抽出する．抽出する際のデータ構造は，lightblue の辞書を修正することを踏まえ，用言の表層形，統語範疇，文の中での用言の開始位置，終了位置の 4 つ組のリストとする．

[1] https://github.com/ku-nlp/KyotoCorpus
[2] https://sites.google.com/site/naisttextcorpus/
[3] https://github.com/ajb129/KeyakiTreebank
[4] https://github.com/DaisukeBekki/lightblue

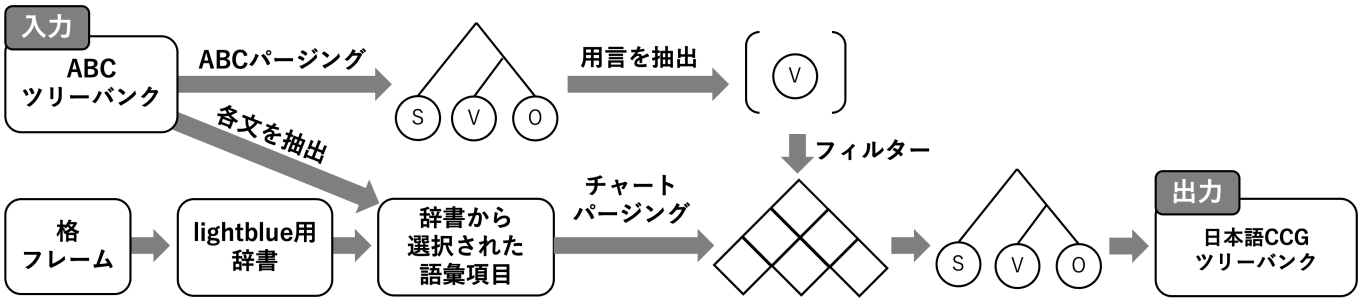

図 3: ABC ツリーバンクのリフォージング

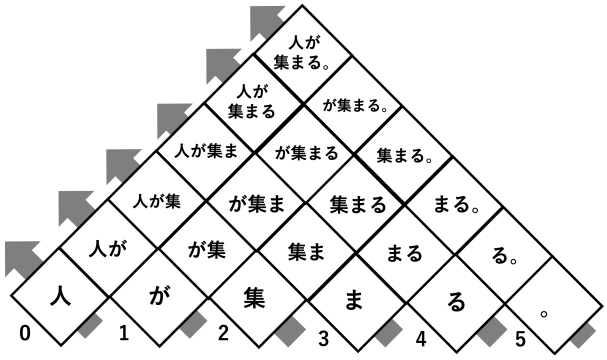

図 4: left-corner chart parsing の実行手順

3.2 lightblue の辞書の書き換え

lightblue 内の `setupLexicon` 関数へ ABC ツリーバンクの文を渡す．`setupLexicon` 関数とは，受け取った文を解析するために必要な語彙項目のリストを返す関数である．ここで獲得した語彙項目リストを用いて left-corner chart parsing を行う．実行手順は図 4 に示す．left-corner chart parsing では，左下から解析を始め，部分文字列すべての可能な組み合わせについてスコアを求め，スコアの高い組み合わせをビーム幅に応じた数だけ出力する．この chart parsing の中に現れる格フレームを持つ用言の語彙項目を ABC ツリーバンクから得られた用言の情報によってフィルターする．これにより，lightblue は従来の lightblue の辞書に登録されている項構造よりも妥当な項構造を用いて統語解析を行うことが可能となる．

3.3 ツリーバンクの再構築

フィルターされたチャートを用いて ABC ツリーバンクの文の統語解析を行う．解析が成功した場合は，項構造が正しく，かつ詳細な統語情報を持った CCG 統語構造が出力される．解析が失敗した場合はエラーが出力される．ABC ツリーバンクのすべての文をパーズし，ツリーバンク形式で出力・保存することで，言語学的に妥当な日本語 CCG ツリーバンクを構築する．

4 評価

ABC ツリーバンクから木構造データをランダムサンプリングし，そのデータから抽出した文に対して，lightblue を用いて統語解析を行う．まず，lightblue でパーズした際に，パーズが成功し CCG 統語構造が出力される確率を調査する．さらに，出力された CCG 統語構造が標準的な日本語 CCG[6] に基づいているかを評価する．

5 おわりに

本稿では，言語学的に妥当な日本語 CCG ツリーバンク構築に向けて，ABC ツリーバンクと lightblue が持つ利点を組み合わせることで，言語学的に妥当かつ詳細な統語素性を持った日本語ツリーバンクを生成する手法を提案した．今後，正しく CCG 統語構造が導出できた文とそのパーズ時間の分析，パーズが失敗した文のエラー分析を行いシステムの改良を行う．

参考文献

[1] 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.

[2] Daisuke Bekki and Hitomi Yanaka. Is Japanese CCGBank empirically correct? A case study of passive and causative constructions. In *Proceedings of Treebanks and Linguistic Theories (TLT) 2023, the workshop in the Georgetown University Round Table on Linguistics 2023 (GURT2023)*, 2023.

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

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

[5] 植松すみれ, 松崎拓也, 花岡洋輝, 宮尾祐介, 美馬秀樹. 統語・意味コーパスの統合と再解釈による大規模な日本語 CCG 文法の開発. 人工知能学会全国大会論文集, Vol. JSAI2013, pp. 4B11–4B11, 2013.

[6] 戸次大介. 日本語文法の形式理論. くろしお出版, 東京, 2010.

[7] 窪田悠介, 峯島宏次, 林則序, 岡野伸哉. 汎用的な範疇文法ツリーバンクの構築. 言語処理学会 第 25 回年次大会発表論文集 (2019 年 3 月), pp. 143–146. 一般社団法人言語処理学会, 2019.

[8] 窪田悠介, 峯島宏次, 林則序, 岡野伸哉. ABC ツリーバンク：学際的な言語研究のための基盤資源. 言語処理学会 第 27 回年次大会 発表論文集 (2021 年 3 月), pp. 1529–1534. 一般社団法人 言語処理学会, 2021.