D2-2

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

言語処理学会第30回年次大会(NLP2024) 2024/03/12

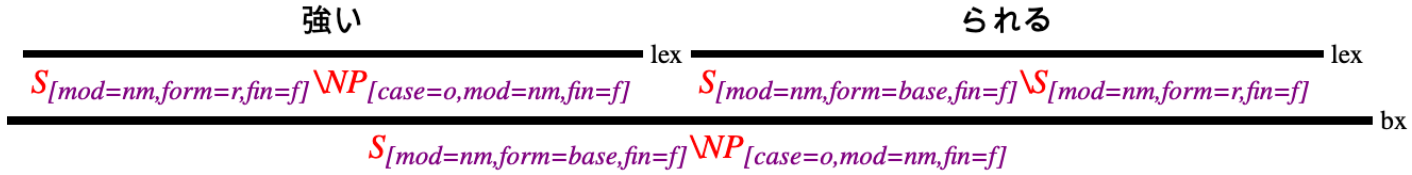
ツリーバンク

大規模なテキストに統語構造が付与されたコーパス

- パーザの学習・評価用データとしても利用される
- 形式統語論に基づいたツリーバンクの例：

  英語
  - Penn Treebank [Marcus+ 1993]（CFG）
  - CCGbank [Hockenmaier and Steedman 2007]（CCG）

  日本語
  - 日本語CCGbank [Uematsu+ 2013]（CCG）
  - ABCツリーバンク [Kubota+ 2019]（ABC文法）



組合せ範疇文法

Combinatory Categorial Grammar (CCG) [Steedman 1996, 2001]

辞書 (lexicon)
語とその統語情報や意味情報を関連付ける

$$\text{Taro} \vdash NP_{3SM}$$
$$\text{runs} \vdash S \setminus NP_{3S}$$

組合せ規則 (combinatory rules)
統語情報、意味情報を並行して組み上げる

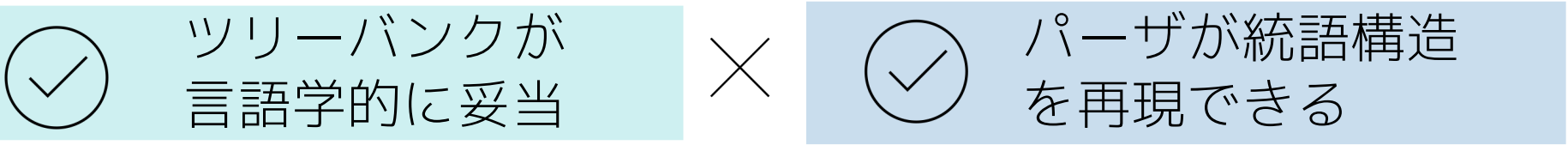
Backward Application (>)

$$Y, \quad X \setminus Y \Rightarrow X$$

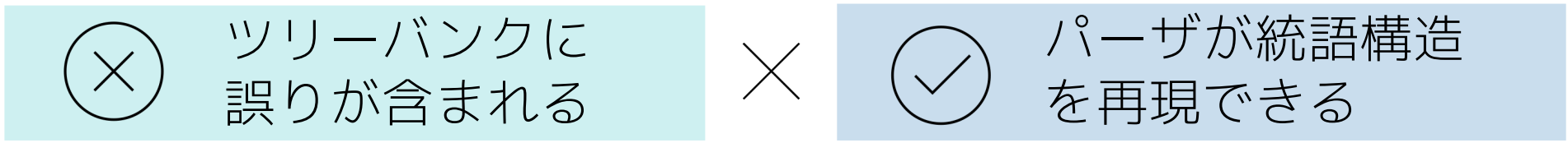
$$\frac{\dfrac{\text{Taro}}{NP_{3SM}} \qquad \dfrac{\text{runs}}{S \setminus NP_{3S}}}{S}\ >$$

CCGパーザと妥当性

- depccg [Yoshikawa+ 2017], jigg [Noji and Miyao 2016] などのCCGパーザが開発されており、ccg2lambdaなどの意味解析・推論システムにも活用されている
- パーザの言語学的な**妥当性**　≠　パーザの**精度**



→ 言語学的に妥当な統語構造が出力できる



→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

日本語CCGbank [Uematsu+ 2013]

- 日本語CCGパーザの学習・評価データとして利用されている
- 受身・使役の文に誤りが含まれていることが指摘されている
  [Bekki and Yanaka 2023]

日本語CCGbankの分析

$$\dfrac{\dfrac{\textbf{太郎が}}{NP_{ga}} \quad \dfrac{\dfrac{\textbf{次郎に}}{NP_{ni}} \quad \dfrac{\dfrac{\textbf{褒めら}}{S\backslash NP_{ga}\backslash NP_{ni}} \quad \dfrac{\textbf{れ}}{S\backslash S}}{S\backslash NP_{ga}\backslash NP_{ni}}<B_2 \quad \dfrac{\textbf{た}}{S\backslash S}}{S\backslash NP_{ga}\backslash NP_{ni}}<B_2}{S\backslash NP_{ga}}< \over S}<$$

日本語CCGbank [Uematsu+ 2013]

- 日本語CCGパーザの学習・評価データとして利用されている
- 受身・使役の文に誤りが含まれていることが指摘されている
  [Bekki and Yanaka 2023]

言語学的に妥当なCCG木

$$\dfrac{\dfrac{\textbf{太郎が}}{T/(T\backslash NP_{ga})}\ \dfrac{\dfrac{\dfrac{\textbf{次郎に}}{T/(T\backslash NP_{ni})}\ \dfrac{\dfrac{\textbf{褒めら}}{S\backslash NP_{ga}\backslash NP_{o}}\ \dfrac{\textbf{れ}}{S\backslash NP_{ga}\backslash NP_{ni}\backslash(S\backslash NP_{ga}\backslash NP_{ni|o})}<}{S\backslash NP_{ga}\backslash NP_{ni}}\ \dfrac{\textbf{た}}{S\backslash S}<}{S\backslash NP_{ga}\backslash NP_{ni}}<B^2}{S\backslash NP_{ga}}>}{S}>$$

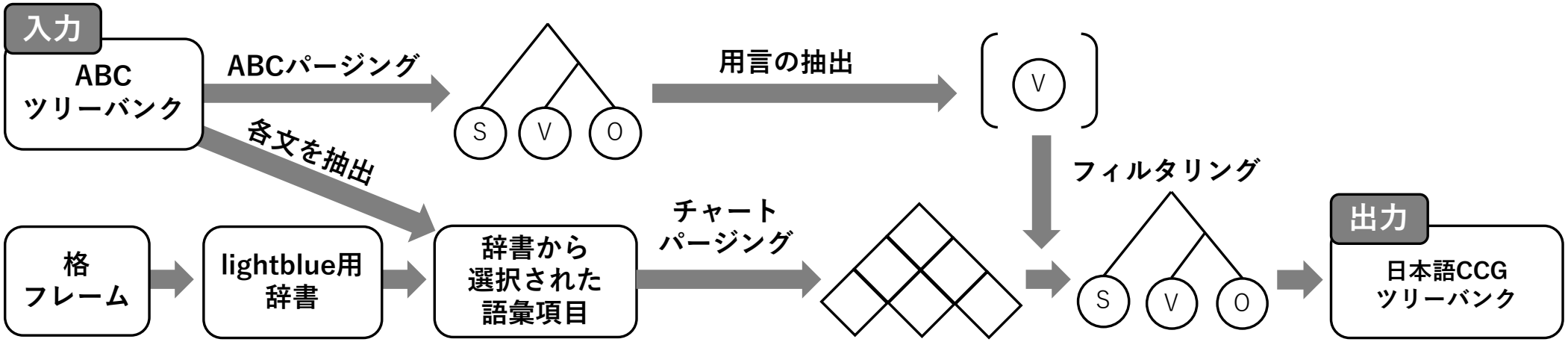
[富田ら 2023]

- 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

ABCツリーバンク [Kubota 2019]
人手によって**項構造が正確**に記述されている

×

lightblue [Bekki and Kawazoe 2016]
詳細な統語情報を与えることができる



本研究の目的

- 富田ら (2023) のアルゴリズムの改良
  - 複合動詞を含む文の分析
- 改良したアルゴリズムに基づいて言語学的に妥当な日本語CCGツリーバンク **lightblue CCGbank** を構築
- lightblue CCGbank の統語構造と意味表示についての評価・考察・エラー分析を行う

lightblue の基本仕様 [1/3]

1 格フレームを元に作られた**辞書約8万語**とCCGの組合せ規則に基づいてCCG統語解析を行う

lightblue の基本仕様 [2/3]

2 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、Left-corner chart parsing を行う

- 部分統語構造を子の統語構造から作る際に、統語範疇、意味表示などが計算される

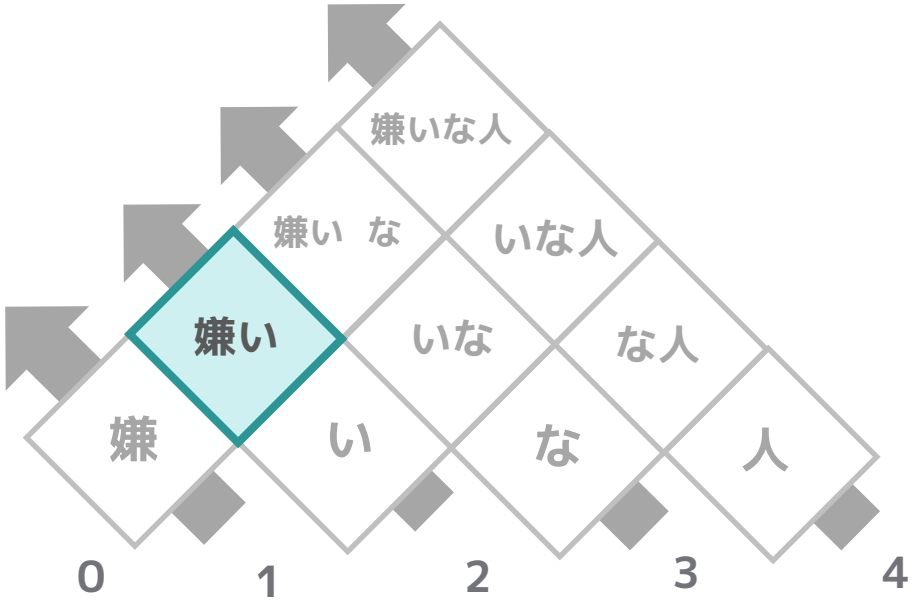
例：(0,2)の語彙項目の一部

(“嫌い/きらい”,

$S_{[n:da|n:na|n:ni|+][nstem]} \setminus NP_{ga}$)

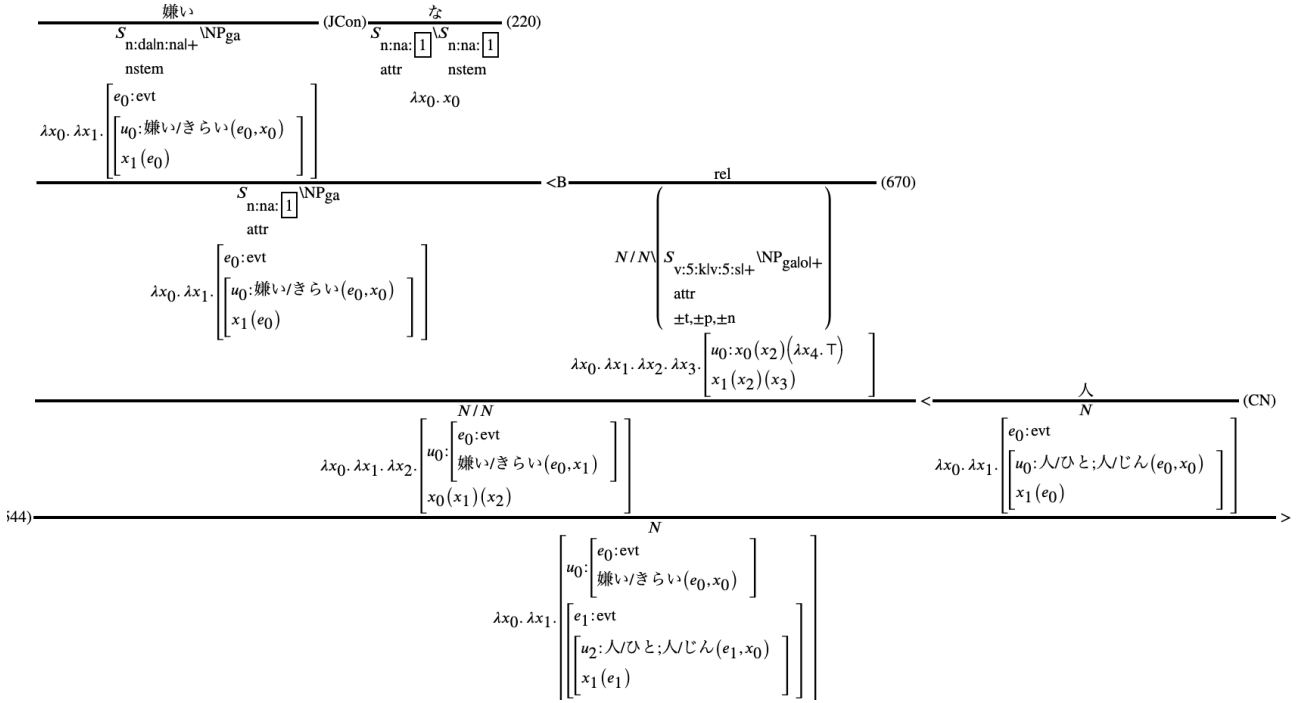
(“嫌う/きらう/ガオ”,

$S_{[v:5:w<1>][cont|mod:m]} \setminus NP_{ga} \setminus NP_{o}$)



lightblue の基本仕様 [3/3]

3 CCG の統語構造と依存型意味論(Dependent Type Semantics, DTS) [Bekki and Mineshima 2017] に基づく意味表示を解析結果として出力する。



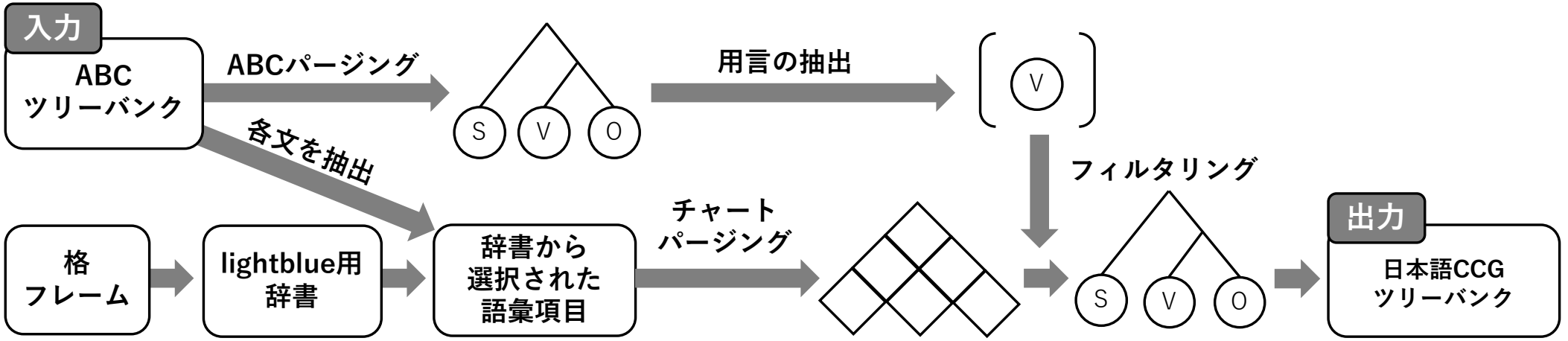
富田らによる提案手法 (再掲)

- 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

ABCツリーバンク [Kubota 2019]
人手によって**項構造が正確**に記述されている

×

lightblue [Bekki and Kawazoe 2016]
詳細な統語情報を与えることができる



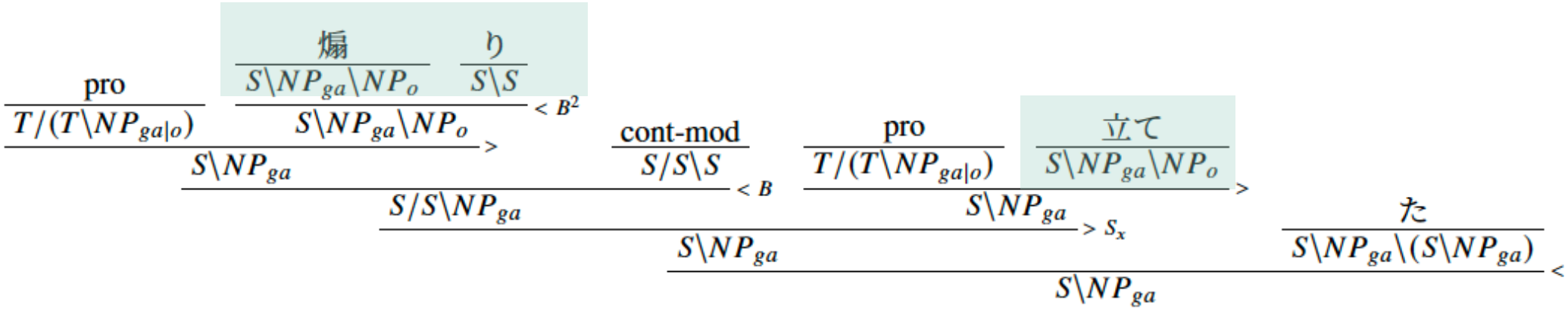
アルゴリズムの改良 [1/4]

課題

複合動詞を含む文において、２つの動詞が分割して解析されてしまう

入力文

急進派の発言は、国民の不安を煽り立てた



複合動詞の分類

- 複合動詞は２種類に分類できる
    1. まとめて一語とされるもの
        例：「煽り立てる」
    2. 動詞＋動詞性接尾語（補助動詞）の合成
        例：「食べ始める」

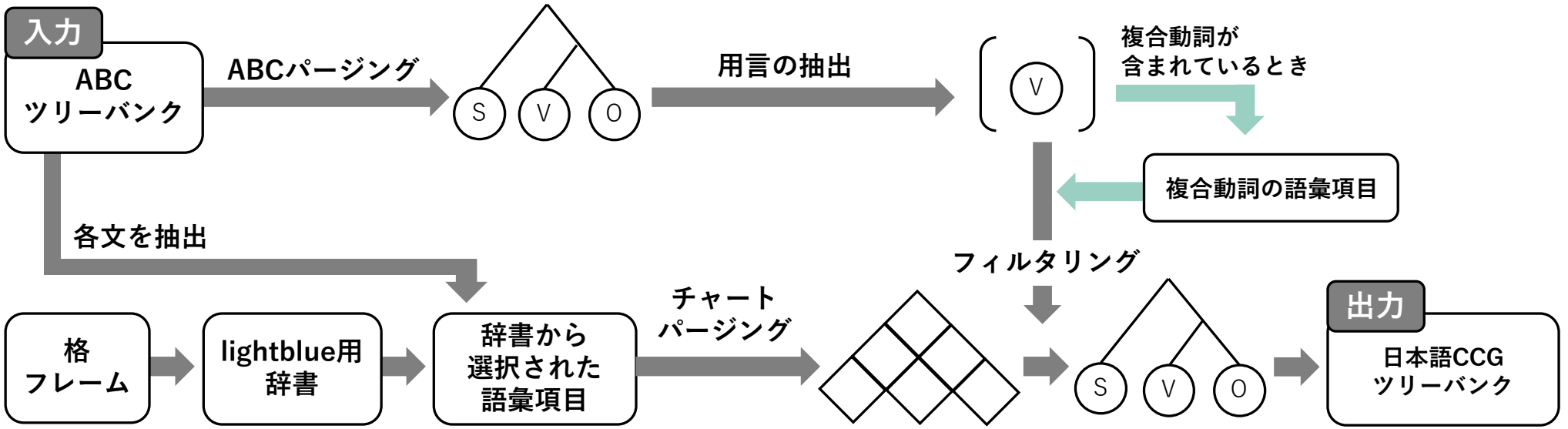
- 今回は全ての複合動詞について、一語となるようにアルゴリズムを改良することを目指す

アルゴリズムの改良 [2/4]

- ABCツリーバンクから抽出した複合動詞のうち、*lightblue* の辞書には登録されていない複合動詞が存在する
  → 誤った解析につながる

- *lightblue* に複合動詞の語彙項目を追加してから統語解析を行うようにアルゴリズムを変更する

アルゴリズムの改良 [3/4]



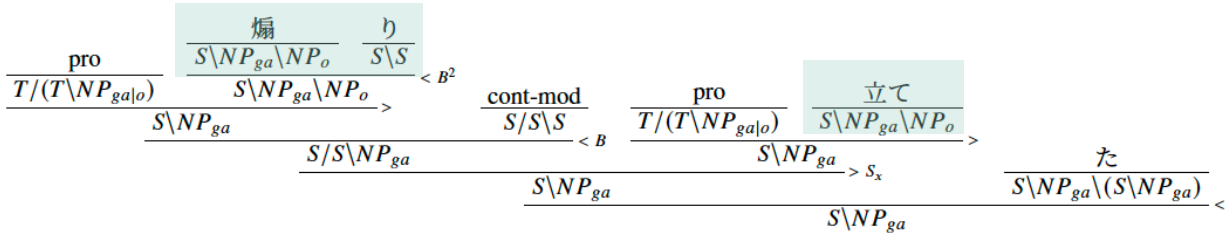
0. あらかじめABCツリーバンクに登場する複合動詞のリストを用意する
1. ABCツリーバンクから抽出した用言のリストの中に複合動詞のリストに含まれている用言があるか判定する
2. *lightblue*で、複合動詞の語彙項目を生成し、語彙項目をフィルタリングする
3. フィルタリングされた語彙項目を用いてチャートパージングを行う

アルゴリズムの改良 [4/4]

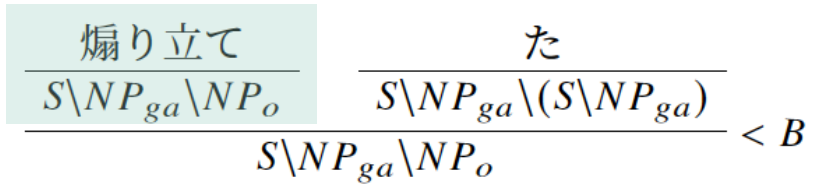
入力文

急進派の発言は、国民の不安を煽り立てた

改良前



改良後



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

lightblue CCGbankの構築

- ABCツリーバンクの14ジャンルから最大5ファイルずつランダムに抽出し、**計15,137文からなるツリーバンク**を構築した
- 1文が50文字を超える文は*lightblue*のパーズに時間がかかるため除外

ジャンル	50字以上の文	文の数
aozora	590	1183
bible	220	1140
book_expert	4	41
dict_lexicon	4	2636
diet_kaigiroku	112	374
fiction	44	877
law	128	209
misc	59	276

ジャンル	50字以上の文	文の数
news	103	340
non-fiction	87	126
spoken	11	559
ted_talk	54	551
text-book	10	4870
wikipeda	51	171
Total	1482	13653

評価

- 「言語学的妥当性」の評価は、出力のCCG統語構造とDTS意味表示を人手で確認する必要がある

- 人手での評価は、コストがかかる作業であるため、評価用データとして、各ジャンルから4文ずつランダムサンプリングし、56文に対して人手での評価を行った

評価結果

- 統語構造、意味表示ともに正しいCCG木は56文中**19文（33%）**

- エラーの中には、lightblue由来のエラー、ABCツリーバンク由来のエラー、追加したアルゴリズム由来のエラーがある

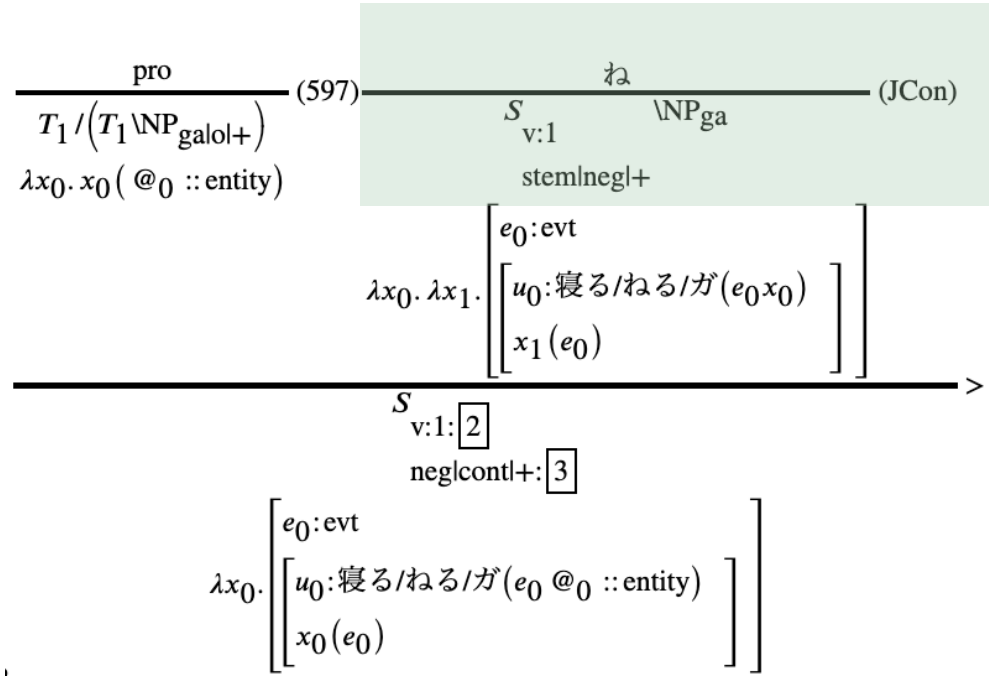
評価軸		エラー文の数
統語構造に関するエラー	未登録語・統語範疇のエラー	8
	複合動詞のエラー	4
	その他エラー	30
意味表示に関するエラー		7

未登録の機能語に由来するエラー

入力文

国際関係の仕事に今ついてるのね

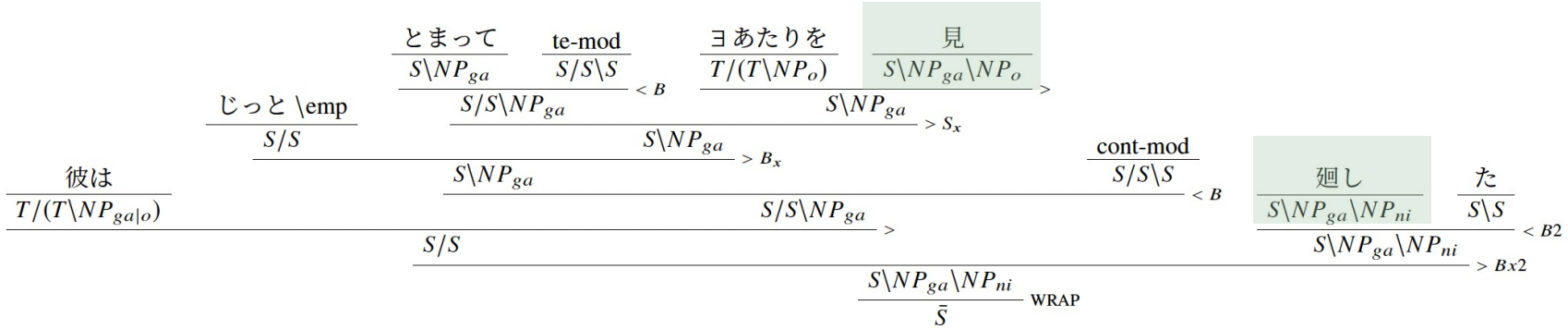
終助詞の「ね」が*lightblue*に登録されていない
→*lightblue*に終助詞「ね」を追加し、適切な制限を設ける必要がある



複合動詞に由来するエラー

入力文

彼はじっととまって、あたりを見廻した。

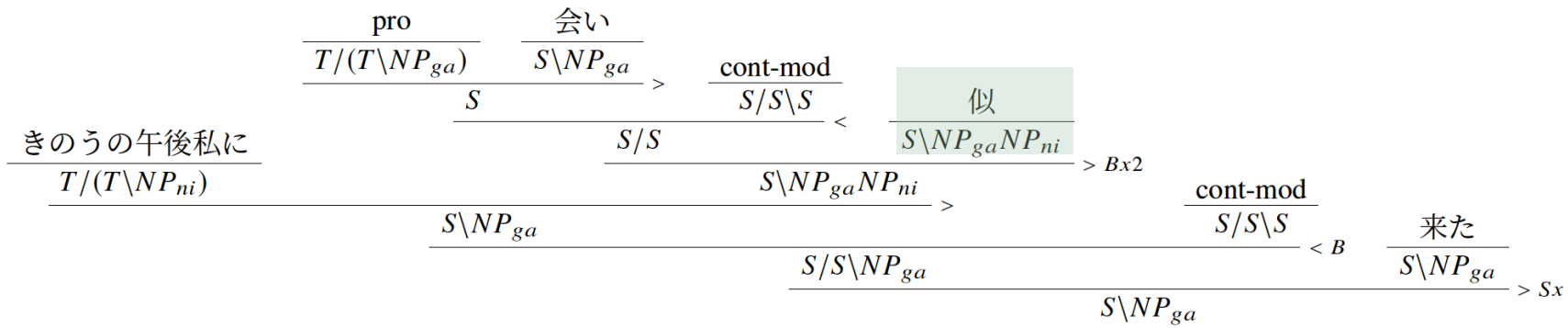


- 「見」と「廻し」が結合されるよりも前に、「見」の前にある「あたりを」というヲ格名詞と結合してしまうことで引き起こされている
- 語彙項目を新しく追加することは実現できたが、新しく作成した語彙項目の優先順位が低くなってしまっている

その他統語構造に関するエラー

入力文

太郎は、きのうの午後私に会いに来た。



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。

まとめ

- 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、複合動詞の扱いを中心にアルゴリズムの改良をおこなっている
- 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を構築した
- ツリーバンクのCCG統語構造とDTS意味表示について評価を行った
  - 改良したアルゴリズムは現時点ではすべての複合動詞について有効ではない

今後の展望

- ツリーバンク構築アルゴリズムのさらなる改良
- *lightblue* の辞書の拡張を行い、ツリーバンクの妥当性を向上させる

近日公開予定

lightblue CCGbank

参考文献

1. Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of En- glish: 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 ex- tracted from the Penn Treebank. Computational Lin- guistics, 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 Con- ference, 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 Lin- guistics (Volume 1: Long Papers), pp. 1042–1051, Sofia, Bulgaria, August 2013. Association for Computa- tional Linguistics.
7. Hiroshi Noji and Yusuke Miyao. Jigg: A framework for an easy natural language processing pipeline. In Pro- ceedings 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 (Vol- ume 1: Long Papers), pp. 277–287, Vancouver, Canada, July 2017. Association for Computational Linguistics.
9. Pascual Mart´ınez-Go´mez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. ccg2lambda: A composi- tional 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 CCG- Bank empirically correct? A case study of passive and causative constructions. In Proceedings of the 21st International Workshop on Treebanks and Linguis- tic Theories (TLT, GURT/SyntaxFest 2023), pp. 32–36, Washington, D.C., March 2023. Association for Compu- tational 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 Com- putational 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 com- pilation 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.