

ABC ツリーバンク：学際的な言語研究のための基盤資源

窪田 悠介
国立国語研究所

kubota@ninjal.ac.jp

林 則序
東京大学

hayashi@phiz.c.u-tokyo.ac.jp

峯島 宏次
慶應義塾大学

minesima@abelard.flet.keio.ac.jp

岡野 伸哉
東京大学

okano@phiz.c.u-tokyo.ac.jp

1 はじめに

NLP と理論言語学の融合的研究に活用可能な日本語の範疇文法ツリーバンクである ABC ツリーバンクを構築した。ツリーバンクのデータ、構築のために開発したツール群、マニュアルは、<https://github.com/ABCTreebank/> で公開する。本論文では、主要な言語現象のアノテーションを概説し、活用事例として、このツリーバンクを学習データとして統語パーサ (CCG パーサ) を訓練し、論理的意味表示に基づいてパーサの性能を評価する実験を試行的に行った結果について報告する。¹⁾

2 ABC 文法の概要

ABC ツリーバンクで用いられているのは ABC 文法という範疇文法の種類である。これは組み合わせ範疇文法 (CCG; [14, 18]) やタイプ論理文法 (TLG; [12, 11, 7]) といった特定の範疇文法の流派へのコミットメントを避けながらも、それらの文法への変換の土台となりうる、汎用性を重視した中間的枠組みである。以下、ABC 文法の基本的な構成要素であるカテゴリと素性、および種々の規則を示す。

2.1 カテゴリと素性

ABC 文法で用いられている基本カテゴリ (atomic category) のうち、代表的なものを表 1 に挙げる。²⁾

カテゴリには素性が表示されることがある。³⁾ 素性は小文字で始め、大文字のカテゴリに続けて表記

1) ABC ツリーバンクは、けやきツリーバンク (<https://github.com/ajb129/KeyakiTreebank>) [4] からの変換によって構築された。構築の詳細については [19, 8] を参照。

2) S, CP はそれぞれ、元コーパスのけやきツリーバンクでの IP, CP に対応し、生成文法での区分とは必ずしも一致しない。

3) 素性による区別は、けやきツリーバンクにおける拡張タグを用いた区別におおむね準ずる。

カテゴリ	名称	代表的な素性
CP	節	f, t, t-sbj, q, q-sbj, x
N	名詞	
NP	名詞句	q
PP	後置詞句	s, s2, o1, o2
Q	量化詞	
S	節	a, e, imp, m, nml, rel, smc, sub

表 1 代表的な基本カテゴリ

する。以下では代表的な素性をカテゴリと合わせて列挙する。

文法役割を示すもの PPs (主語), PPs2 (第二主語), PPO1 (第一目的語), PPO2 (第二目的語), CPt-sbj (補部節主語), CPq-sbj (疑問節主語)⁴⁾

節の種類を示すもの Sa (連用節), Se (空所なし名詞修飾節), Simp (命令節), Sm (主節), Snml (名詞化節), Srel (関係節), Ssmc (小節), Ssub (準主節), CPF (終助詞節), CPq (疑問節) CPt (補部節), CPx (感嘆節)

量的であることを示すもの NPq

一般に範疇文法では、X, Y がカテゴリであるならば、X/Y, Y\X もカテゴリ (関数的カテゴリ) である。例えば、自動詞は PPs\S, 他動詞は PPO1\ (PPs\S) というカテゴリを付与される。

2.2 統語規則

ABC 文法という名前には、関数適用の規則のみをもつ AB 文法 [1, 3] に、関数合成 (Function Composition) の規則を加えたもの、という由来がある。したがってこの 2 つは基本的な規則として置か

4) CP のうち、文法役割が素性で明示されるのは主語の場合のみである。

れている。図 1 にこれを示す。⁵⁾

関数適用	関数合成
$A/B \ B \Rightarrow A$	$A/B \ (B/C)/\$ \Rightarrow (A/C)/\$$
$B \ B \setminus A \Rightarrow A$	$\$(C \setminus B) \ B \setminus A \Rightarrow \$(C \setminus A)$

図 1 基本的な規則

上に示した binary 規則だけでなく、ある 1 つのカテゴリを別のカテゴリに変換する、unary 規則も存在する。これは日本語を分析するにあたっての必要性から用いられるもので、上記の一般的な規則とは異なり、適用可能なカテゴリ・素性が限られている。unary 規則は以下のものに大別される。(i) 名詞の項化・述語化；(ii) 関係節による名詞修飾；(iii) 連用節化・副詞化；(iv) かき混ぜ。以下、規則において素性の指定がない場合は、カテゴリさえ合致していればそれが適用可能であるということを意味する。

2.2.1 名詞の項化・述語化

名詞 N はまず名詞句 NP を投射した上で、さらに述語または項としてのカテゴリを投射する。以下、関連する unary 規則を簡単に説明する(表 2 参照)。

名詞の名詞句化 本ツリーバンクでは、基本カテゴリとして N と NP の区別を行い、N から NP へは、つねに 1 本の unary branch が伸びる。

名詞句の述語化 コピュラが後続する名詞句はそれ自体が述語と見なされ、S で終わるカテゴリとなる(コピュラは S \ S のようなカテゴリを与えられ、関数合成で述語名詞句と結びつく)。

名詞句の後置詞句化 表層で格助詞が落ちる名詞句も、格助詞のある場合と同じカテゴリとして扱うため unary 規則で PP を投射する。

規則の種類	規則
名詞の名詞句化	$N \Rightarrow NP$
名詞句の述語化	$NP \Rightarrow PPs \setminus S$
名詞句の後置詞句化	$NP \Rightarrow PPs, NP \Rightarrow PPs \setminus S$

表 2 名詞の項化・述語の規則(抜粋)

2.2.2 関係節による名詞修飾

日本語には英語のような明示的な関係代名詞が存在しないため、表 3 に代表される規則を用いて関係節(Srel)を名詞を修飾するカテゴリ(N/N)へと変換する。

5) $X/\$$ は $X, X/Y, (X/Y)/Z, \dots$ のいずれかを、 $\$ \setminus X$ は $X, Y \setminus X, Z \setminus (Y \setminus X), \dots$ のいずれかを表す。

規則の種類	規則
主語に対する内の関係	$PPs \setminus Srel \Rightarrow N/N$
第一目的語に対する内の関係	$PPo1 \setminus Srel \Rightarrow N/N$
付加句に対する内の関係	$Srel \Rightarrow N/N$

表 3 関係節による名詞修飾の規則(抜粋)

2.2.3 連用節化・副詞化

ABC ツリーバンクは、変換元であるけやきツリーバンクの仕様を引き継ぎ、動詞の連用形と終止形をカテゴリの上で区別しない。一方で、両ツリーバンクは連用節に特別のカテゴリを与えている(けやきツリーバンクにおける連用節 IP-ADV が ABC ツリーバンクの Sa に対応する)。この情報を利用し、動詞の連用形が「て・で」などの接続助詞を伴わずに連用節をなすケースについて、表 4 上段の規則を導入する。つまり、Sa で終わるカテゴリから、S で終わるカテゴリを修飾するカテゴリへの変換である。

加えて、「今朝」のような副詞的な用法をもつ名詞句についても、同様に S で終わるカテゴリを修飾するカテゴリへの変換規則が必要である(表 4 下段)。

規則の種類	規則
連用節化	$Sa \Rightarrow S/S,$ $PPs \setminus Sa \Rightarrow (PPs \setminus S)/(PPs \setminus S)$
名詞句の副詞化	$NP \Rightarrow S/S,$ $NP \Rightarrow (PPs \setminus S)/(PPs \setminus S)$

表 4 連用節化・副詞化の規則(抜粋)

2.2.4 かき混ぜ

ABC ツリーバンクでは、かき混ぜ(scrambling)を規則で扱う。例えば、「主語-目的語-述語」という語順と、「目的語-主語-述語」という語順に対応して、同じ音形の述語に別個の語彙項目を割り当てのではなく、前者を基本語順として、それ以外の語順は規則適用によって捉える、という扱いをする(表 5)。

規則の種類	規則
主語と第一目的語	$PPo1 \setminus (PPs \setminus S) \Rightarrow PPs \setminus (PPo1 \setminus S)$
主語と第二主語	$PPs2 \setminus (PPs \setminus S) \Rightarrow PPs \setminus (PPs2 \setminus S)$

表 5 かき混ぜの規則(抜粋)

3 意味表示

範疇文法をフレームワークとして用いる ABC ツリーバンクは、統語構造と意味表示との対応に関して明示的な解析を与える。範疇文法の意味パーサ ccg2lambda [9]、および、辞書としての意味テンプレ

レートを用いることで、 λ 計算に基づく意味合成により、ABC ツリーバンクの木から意味表示とその導出木を自動的に得ることができる。意味表示としては一階論理 (FOL), 談話表示構造 (DRS) [6], 高階論理 (HOL) の 3 つの出力形式を用いる。

具体例として、(1) の各意味表示を以下に示す。

- (1) 窓が開けてある。
(Keyaki:13_textbook_djg_gram_terms)

FOL: $\exists x_1 \exists x_2. (\text{窓}(x_2) \wedge \exists e_3. (\text{開ける}(e_3) \wedge \text{arg_s}(e_3, x_1) \wedge \text{arg_o1}(e_3, x_2) \wedge \text{て}(e_3)))$

DRS: $(\{x_1, x_2, e_3\}, \{\text{窓}(x_2), \text{開ける}(e_3), \text{arg_s}(e_3, x_1), \text{arg_o1}(e_3, x_2), \text{て}(e_3)\})$

HOL: ある $(\lambda y \lambda x. \text{開ける}(x, y), \text{窓})$

それぞれの意味表示について簡単に説明する。

FOL FOL の意味表示は Neo-Davidsonian Event Semantics [13] に基づく。ここで、 $\text{arg_s}(e, x)$ と $\text{arg_o1}(e, x)$ はそれぞれ、ABC 文法のカテゴリ PPs と PPO1 に対応する意味役割であり、「イベント e の主体は個体 x である」「 e の対象は x 」を表す。(1) では、「(て) ある」を受身に類する述語として解析することで、適切な述語項構造が得られている。⁶⁾

DRS DRS は意味合成によって導出された FOL の意味表示からの自動変換によって得られる。DRS は、談話指示物 (discourse referent) の集合 D と条件の集合 C の対 (D, C) から構成され、図 2 のように Box 表記により見通しよく表すことができる。また 4 節で述べるように、DRS を節形式に変換することで意味表示の評価に用いることができる。

HOL HOL は範疇文法の統語構造に即した意味表現であり、範疇文法の意味合成の過程を関数と項の形で明示的に記述することができる。上の HOL の例では、「(て) ある」は高階述語として扱われ、2 項述語「開ける」を項とし、その項構造を変化させる役割を担う。

ここで使用している論理的な意味表示の特徴として、量化や否定、モーダル表現などのスコープを明示的に表現できることが挙げられる。例えば、以下に示す (2) の各意味表示は、理由の「から」および否定「ない」がもつスコープ関係を捉えたものである。

- (2) 咳がなおらないから、病院に行きます。
(642_textbook_purple_basic)

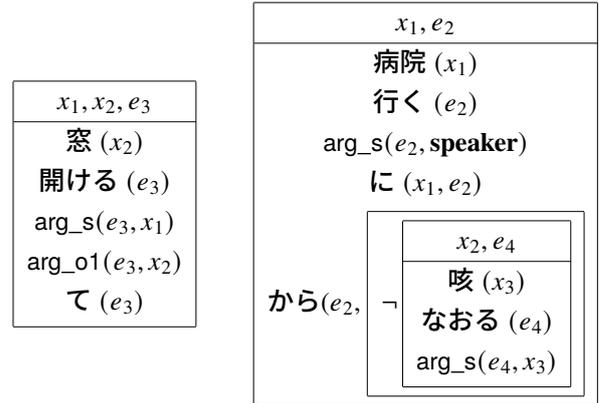


図 2 (1) と (2) の DRS の Box 表記

FOL: $\exists x_1. (\text{病院}(x_1) \wedge \exists e_2. (\text{行く}(e_2) \wedge \text{arg_s}(e_2, \text{speaker}) \wedge \text{に}(e_2, x_1) \wedge \text{から}(e_2, \neg \exists x_3. (\text{咳}(x_3) \wedge \exists e_4. (\text{なおる}(e_4) \wedge \text{arg_s}(e_4, x_3))))))$

DRS: $(\{x_1, e_2\}, \{\text{病院}(x_1), \text{行く}(e_2), \text{arg_s}(e_2, \text{speaker}), \text{に}(e_2, x_1), \text{から}(e_2, \neg (\emptyset, \{\{x_3, e_4\}, \{\text{咳}(x_3), \text{なおる}(e_4), \text{arg_s}(e_4, x_3)\})\}))\})$

HOL: から $(\neg \text{なおる}(\text{咳}), \text{に}(\text{病院}, \lambda x. \text{行く}(x)), \text{speaker})$

現時点で意味表示の導出に用いられる意味テンプレートは、unary 規則に対するもの (46 個), カテゴリに対する意味割り当て (62 個), カテゴリと単語のペアに対する意味割り当て (68 個) からなる。付録 A に (1) と (2) の意味表示を導く ABC ツリーバンクの導出木を示す。

4 統語パーサの構築と評価

ABC ツリーバンクの応用として日本語統語パーサの構築が考えられる。以下、この途中経過を報告する。係り受け以上の細かい分析を行うパーサに関しては、木構造の些末な差異などが性能評価に大きな影響を及ぼしがちであり、適切な評価指標は何かという問題がつきまとう。一つの解決策として、パーサの性能評価を構文木のレベルで行うのではなく、構文木に基づいて得られる意味表示 (3 節参照) のレベルで行う方法が考えられる。以下、この手順で行った性能評価の結果を報告する。

まず、ABC ツリーバンクの木から生成された意味表示の妥当性を人手でチェックし、評価用の GOLD データ (223 件) を作成した。言語現象の多様性を確保するため、言語学のテキストの例文を集めたもの (textbook, 133 例) を中心とし、その他に書籍 (book, 11 例), 辞書 (dict, 17 例), フィクション (fiction, 26 例), ニュース (news, 8 例), ノンフィクション (nonfiction, 13 例), テッドトーク (ted,

6) 使役・受身やコントロール述語などの文末述語の扱いについて詳しくは [19, 8] を参照。

15 例) という複数のジャンルの文を含めた。

今回は, [17] と同様の手法によって, 範疇と木の依存関係の両方を表現ベクトルの中に含んだ bi-LSTM モデルを, AllenNLP [5] を用いて訓練した。ABC ツリーバンクの全データ (44,750 文) から上記の評価用データ (223 文) を除き, さらに二分木でないもの⁷⁾を排除した 21,456 文のうち, 訓練用に 17,137 文, 訓練のステップごとの評価のために 4,319 文を分けて, 訓練を行った。⁸⁾

ここで学習したパーサの評価のため, パーサの出力の木からテンプレートにより意味表示を導出し, それを DRS の節形式に変換する。節形式は, 節の一致率を求める評価手法 (COUNTER [16]) によって比較することが可能である。⁹⁾(2) の節形式を図 3 に示す。

b1 REF x1	b3 REF x3
b1 REF e2	b3 REF e4
b1 病院 x1	b3 咳 x3
b1 行く e2	b3 なおる e4
b1 arg_s e2 speaker	b3 arg_s e4 x3
b1 に e2 x1	b1 から e2 b2
b2 NOT b3	

図 3 (2) の DRS から得られた節形式

ABC パーサで評価用データの 223 文を解析¹⁰⁾し, 出力の導出木から DRS の節形式を生成した。COUNTER による GOLD の節形式との一致率は, precision の平均が 77.4%, recall の平均が 76.4%, F1 の平均が 76.7% であった。意味表示が出力されず解析エラーとなったものは, 10 文であり, 残りの 213 文に対しては well-formed な意味表示が出力された。

以下では一致率が低かった文に関するエラー分析の結果を簡単に報告する。問題点は主に, 係り受けや連体節の種類 (「内の関係」対「外の関係」) など, 統語解析のみでは解決が困難な問題に関するパーサの誤解析と, 変換元コーパスのけやきツリーバンクの仕様に依存する部分でのエラーに大別できる。

パーサの誤解析に関しては, (3) のような例 (prec/rec/F1 = 42.9/45.0/43.9) で「内の関係」の連体節が誤って「外の関係」として解析されるケースや,

7) 例えば, 複数文の埋め込み multi-sentence や, 断片的な文 FRAG を範疇とする部分木を含む文が該当する。

8) 結果とパラメータは, https://github.com/ABCTreebank/ABCTreebank/tree/master/papers/202103_NLP27 で公開予定。

9) これは, Abstract Meaning Representation (AMR [2]) の評価で使われている SMATCH をスコップ付きの意味表現に拡張した評価法である。詳細は [16] を参照。

10) それらの文の形態素解析は, もとの ABC ツリーバンクにおける解析をそのまま使用した。

(4) のような例 (50.0/45.5/47.6) で動詞の項の係り受け先が正しくないケースが目立った (「教える」に係るべき「人に」が誤って「獲る」に係っている)。

(3) ... 泳ぎをおぼえるチャンスもなかった。
(7_book_excerpt-18;JP)

(4) 人に魚を獲ることを教えれば, ...
(28_nonfiction_ANC__110CYL067;JP)

ツリーバンクの仕様に依存する問題の代表例としては, 元コーパスのけやきツリーバンクのアノテーションに従って決定される, 動詞の項構造に関するものがある。例えば存在動詞「ある」に対しては, 「に/が」または「が/が」の格配列に応じて, 以下の二つのカテゴリが割り振られる。

(5) a. 私が意欲がある PPs\PPs2\S
b. 私に意欲がある PPo1\PPs\S

以下の (6) のように外の関係の連体節で表層で格助詞が落ちる場合, 「ある」に関して (5a) と (5b) のどちらの語彙項目を使うかの曖昧性がある。

(6) 意欲ある住民 (43_news_KAHOKU_33)

この曖昧性は本来は意味解釈のレベルでは捨象されるべきものだが, 現状では動詞の項構造を示すレベルの不一致による解析の失敗と分類されている。

5 おわりに

[8] で詳述した文末述語の扱いに関しては, 構造が単純な文についてはパーサの解析結果に基づいて概ね正しい意味表示が得られることが確認できた。言語学的知見を適切に反映させた言語資源による end-to-end の解析システムを組むことができることを示した点において, 一定程度所期の目的を達成したと考えられる。今後の課題としては, 4 節で議論した誤解析への対応, 関連研究である日本語 CCGBank [15] に基づく統語・意味パーサ [10] との比較, また, 汎用性を重視したツリーバンクであるという特徴を活かして, CCG では扱いが煩雑になる量化現象を, タイプ論理文法の手法で扱うことなどに取り組みたい。

謝辞 吉川将司氏 (東北大) のサポートに感謝します。この研究は JSPS 科研費 18K00523, 国立国語研究所共同研究プロジェクト「対照言語学的観点から見た日本語の音声と文法」の助成を受けている。

参考文献

- [1] Kazimierz Ajdukiewicz. Die syntaktische Konnexität. In Storrs McCall, editor, *Polish Logic 1920–1939*, pp. 207–231. Oxford University Press, Oxford, 1935. Translated from *Studia Philosophica*, 1, 1–27.
- [2] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 178–186, 2013.
- [3] Yehoshua Bar-Hillel. A quasi-arithmetic notation for syntactic descriptions. *Language*, Vol. 29, pp. 47–58, 1953.
- [4] Alastair Butler, Kei Yoshimoto, Shota Hiyama, Stephen Wright Horn, Iku Nagasaki, and Ai Kubota. The Keyaki Treebank Parsed Corpus, Version 1.1. <http://www.compling.jp/keyaki/>, 2018. Last accessed: 2021.01.01.
- [5] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pp. 1–6, 2018.
- [6] Hans Kamp and Uwe Reyle. *From Discourse to Logic*. Kluwer Academic Publishers, Dordrecht, 1993.
- [7] Yusuke Kubota and Robert Levine. *Type-Logical Syntax*. MIT Press, Cambridge, Mass., 2020.
- [8] Yusuke Kubota, Koji Mineshima, Noritsugu Hayashi, and Shinya Okano. Development of a general-purpose categorial grammar treebank. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 5195–5201, 2020.
- [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, 2016.
- [10] Koji Mineshima, Ribeka Tanaka, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. Building compositional semantics and higher-order inference system for a wide-coverage Japanese CCG parser. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2236–2242.
- [11] Michael Moortgat. Categorial Type Logics. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pp. 93–177. Elsevier, Amsterdam, 1997.
- [12] Glyn Morrill. *Type Logical Grammar: Categorial Logic of Signs*. Kluwer, Dordrecht, 1994.
- [13] Terence Parsons. *Events in the Semantics of English*. MIT Press, Cambridge, Mass., 1990.
- [14] Mark Steedman. *The Syntactic Process*. MIT Press, Cambridge, Mass., 2000.
- [15] Sumire Uematsu, Takuya Matsuzaki, Hiroki Hanaoka, Yusuke Miyao, and Hideki Mima. Integrating multiple dependency corpora for inducing wide-coverage Japanese CCG resources. *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 14, No. 1, pp. 1–24, 2015.
- [16] Rik van Noord, Lasha Abzianidze, Hessel Haagsma, and Johan Bos. Evaluating scoped meaning representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- [17] Masashi Yoshikawa, Hiroshi Noji, and Yuji Matsumoto. A* CCG parsing with a supertag and dependency factored model. *Journal of Natural Language Processing*, Vol. 26, No. 1, pp. 83–119, 2019.
- [18] 戸次大介. 日本語文法の形式理論. くろしお出版, 東京, 2010.
- [19] 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, pp. 143–146, 2019.

A 付録

3節で例示した(1)と(2)の文の意味表示(FOL, DRS, HOL)に対する導出木を示す。これは3節で述べた意味テンプレートに基づいて、ccg2lambdaを用いてABCツリーバンクの木から生成したものである。なお、ABCツリーバンクでは名詞句のカテゴリとしてNとNPを区別しているが(2.2.1節を参照)、現在のところ、意味合成においてNPとNの意味割り当ては区別されていない。

(1) 「窓が開けてある」(Keyaki:13_textbook_djg_gram_terms)の導出木

13_textbook_djg_gram_terms;b_page_2;JP: 窓が開けてある。

FOL-Normalized: exists x1 x2.(窓(x2) & exists e3.(開ける(e3) & arg_s(e3,x1) & arg_o1(e3,x2) & て(e3)))

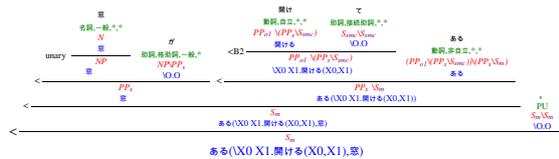
DRS: ([e3,x1,x2],[窓(x2), 開ける(e3), arg_s(e3,x1), arg_o1(e3,x2), て(e3)])

```

x1 x2 e3
窓(x2)
開ける(e3)
arg_s(e3,x1)
arg_o1(e3,x2)
て(e3)
    
```



HOL: ある($\lambda X0 \lambda X1$.開ける($X0, X1$),窓)



(2) 「咳がなおらないから、病院に行きます」(642_textbook_purple_basic)の導出木

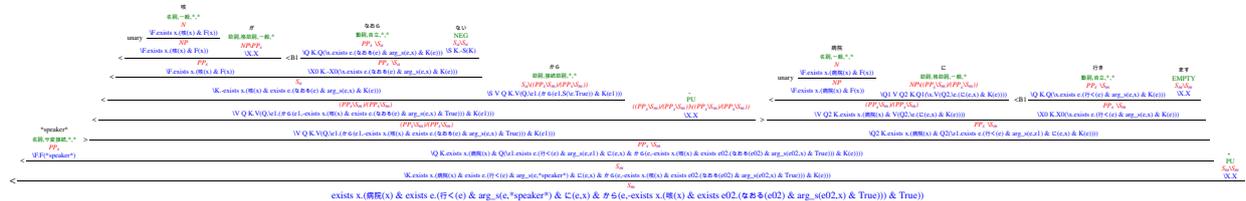
642_textbook_purple_basic;page_217;JP: *speaker* 咳がなおらないから、病院に行きます。

FOL-Normalized: exists x1.(病院(x1) & exists e2.(行く(e2) & arg_s(e2,*speaker*) & に(e2,x1) & から(e2,-exists x3.(咳(x3) & exists e4.(なおる(e4) & arg_s(e4,x3))))))

DRS: ([e2,x1],[病院(x1), 行く(e2), arg_s(e2,*speaker*), に(e2,x1), から(e2,[[-[e4,x3],[咳(x3), なおる(e4), arg_s(e4,x3)]])])])

```

x1 e2
病院(x1)
行く(e2)
arg_s(e2,*speaker*)
に(e2,x1)
から(e2, [
  x3 e4
 咳(x3)
  なおる(e4)
  arg_s(e4,x3)
])
    
```



HOL: から(-なおる(咳),に(病院, $\lambda X0$.行く($X0$)),*speaker*)

