言語学的に妥当な CCGツリーバンク構築に向けて

2023年3月17 日　言語処理学会第29回年次大会併設ワークショップJLR2023

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

目次

01

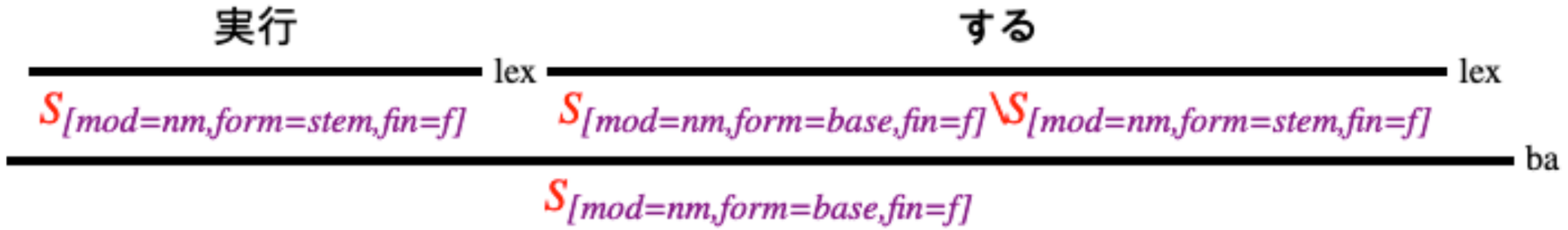
研究概要

ツリーバンク

各文に統語構造が付与されているコーパス

※コーパス : 新聞や本などの内容をテキスト化したデータを大量に集めたデータベース
※統語構造 : 語同士がどのような順番や規則で結合しているのかを表す構造

例： 日本語CCGbank [Uematsu+; 2013]



組合せ範疇文法（Combinatory Categorial Grammar; CCG）

語彙化文法の一種

辞書（統語構造や意味に関わる情報の記述）

例:
Keats ⊢ NP
eats ⊢ $(S \setminus NP) / NP$
apples ⊢ NP

＋

組み合わせ規則

例:
関数適用規則
関数合成規則
型繰り上げ規則

背景と目的

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

CCGパーザの妥当性はCCGツリーバンクの妥当性に依存する

言語学的に妥当なCCGツリーバンクの構築が必要である

→Bekki (2010) に基づいている

02

先行研究

先行研究

01

日本語CCGbank
[Uematsu+ 2013]

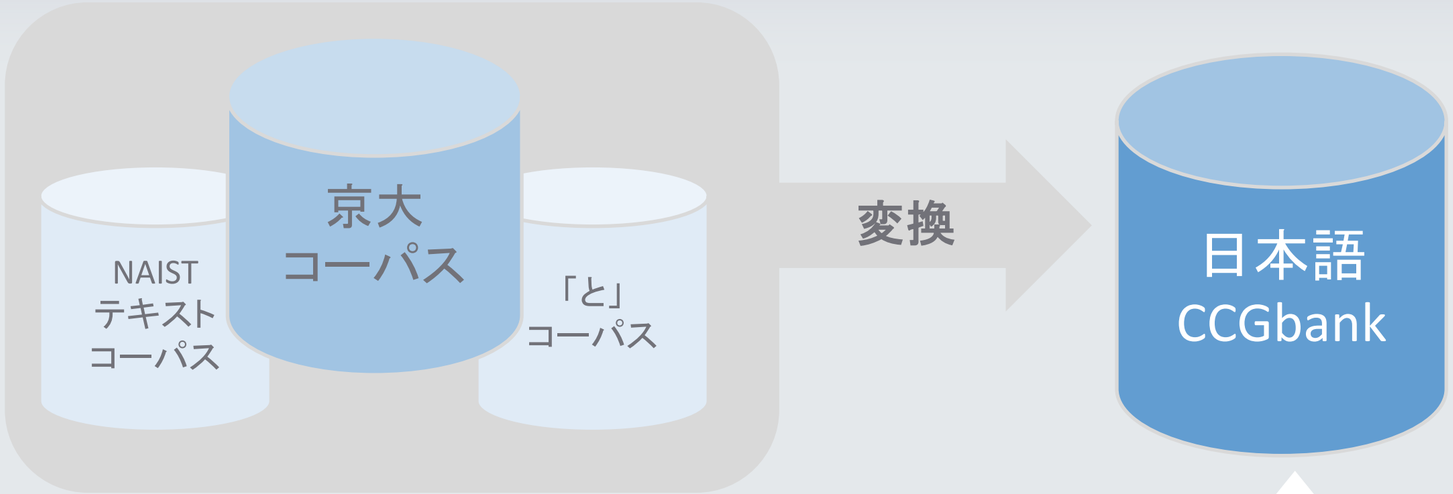
02

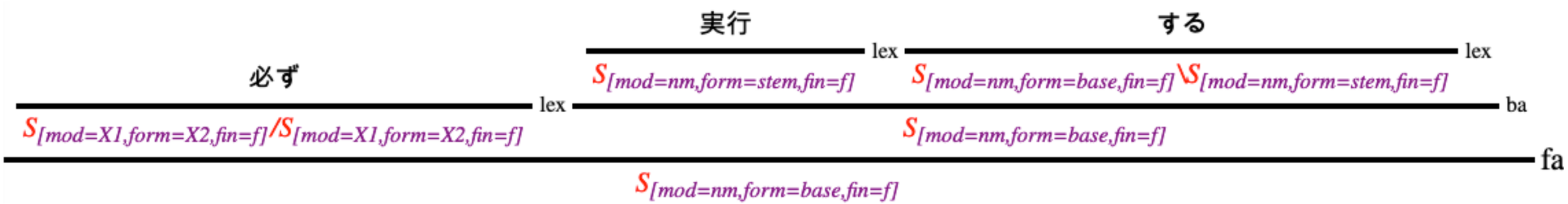
ABC**ツリーバンク**
[Kubota+ 2019]

03

lightblue
[Bekki+ 2016]

日本語CCGbank [Uematsu+, 2013] 1/3





日本語CCGbank [Uematsu+, 2013] 2/3

課題

受身・使役の構文に対して誤った分析がなされている。
[Bekki & Yanaka, 2023]

日本語CCGbankの誤ったCCG木

$$\dfrac{\dfrac{\text{太郎が}}{\mathbf{NP_{ga}}}\quad\dfrac{\dfrac{\text{次郎に}}{\mathbf{NP_{ni}}}\quad\dfrac{\dfrac{\dfrac{\text{褒めら}}{\mathbf{S\backslash NP_{ga}\backslash NP_{o}}}\quad\dfrac{\text{れ}}{\mathbf{S\backslash S}}}{\mathbf{S\backslash NP_{ga}\backslash NP_{o}}}<B^2\quad\dfrac{\text{た}}{\mathbf{S\backslash S}}}{\mathbf{S\backslash NP_{ga}\backslash NP_{o}}}<B^2}{\mathbf{S\backslash NP_{ga}}}<}{\mathbf{S}}>$$

日本語CCGbank [Uematsu+, 2013] 3/3

課題

受身・使役の構文に対して誤った分析がなされている。
[Bekki & Yanaka, 2023]

日本語CCGbankの分析

$$\frac{\text{れ}}{S \backslash S}$$

言語学的に正しいCCG木

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

先行研究

01

日本語CCGbank
[Uematsu+ 2013]

02

ABC**ツリーバンク**
[Kubota+ 2019]

03

lightblue
[Bekki+ 2016]

ABCツリーバンク [Kubota+, 2019] 1/5

- ABC文法を用いたツリーバンク

ABC文法

関数適用のみからなるAB文法＋関数合成(Function Composition)規則

関数適用

$$A / B \quad B \Rightarrow A$$

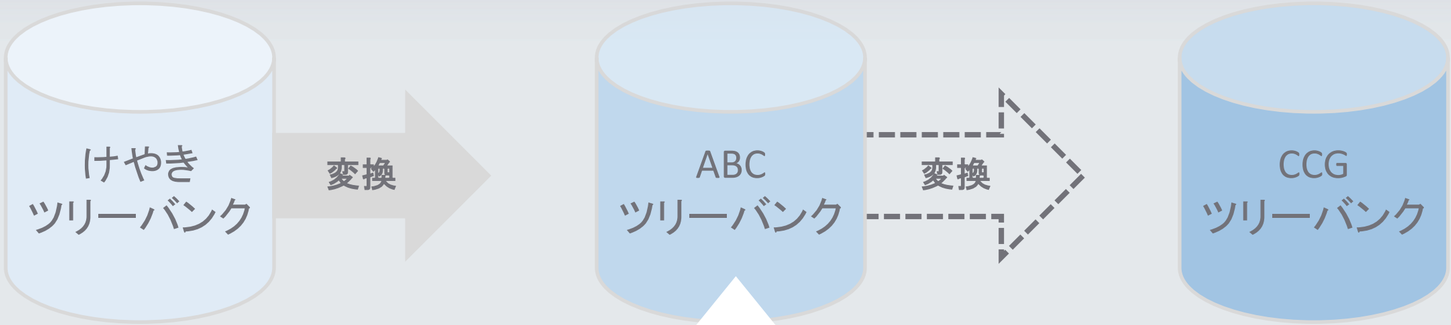
$$B \quad B \setminus A \Rightarrow A$$

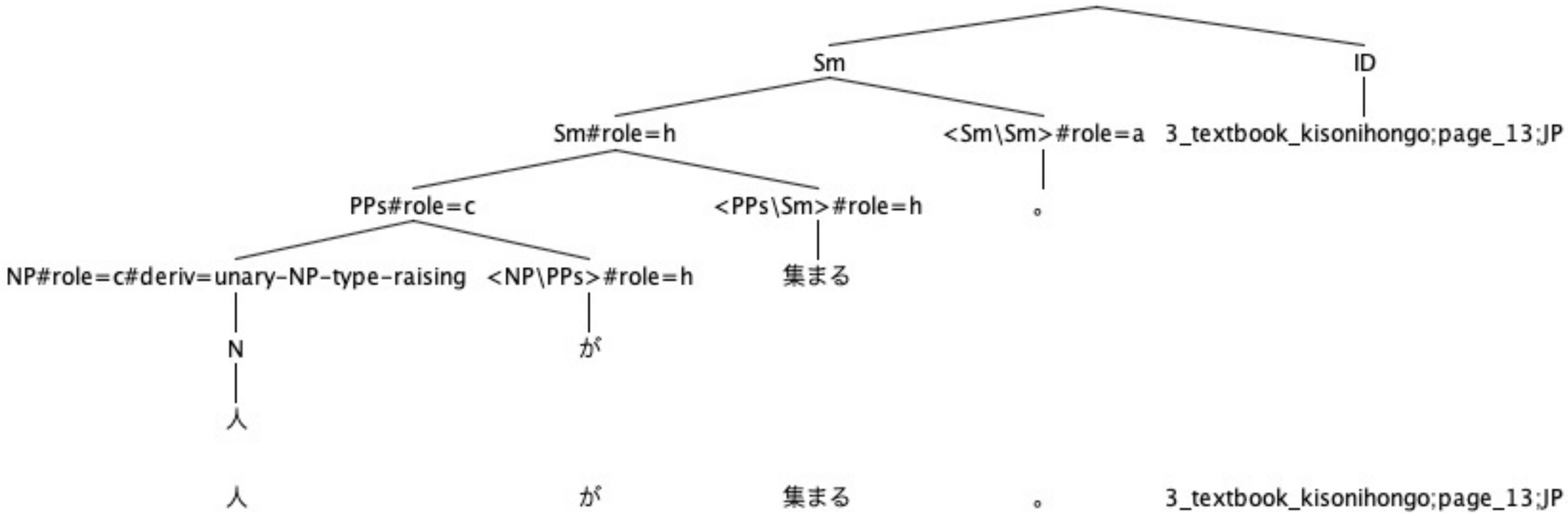
関数合成

$$A / B \quad B / C \Rightarrow A / C$$

$$C \setminus B \quad B \setminus A \Rightarrow C \setminus A$$

ABCツリーバンク [Kubota+, 2019] 2/5

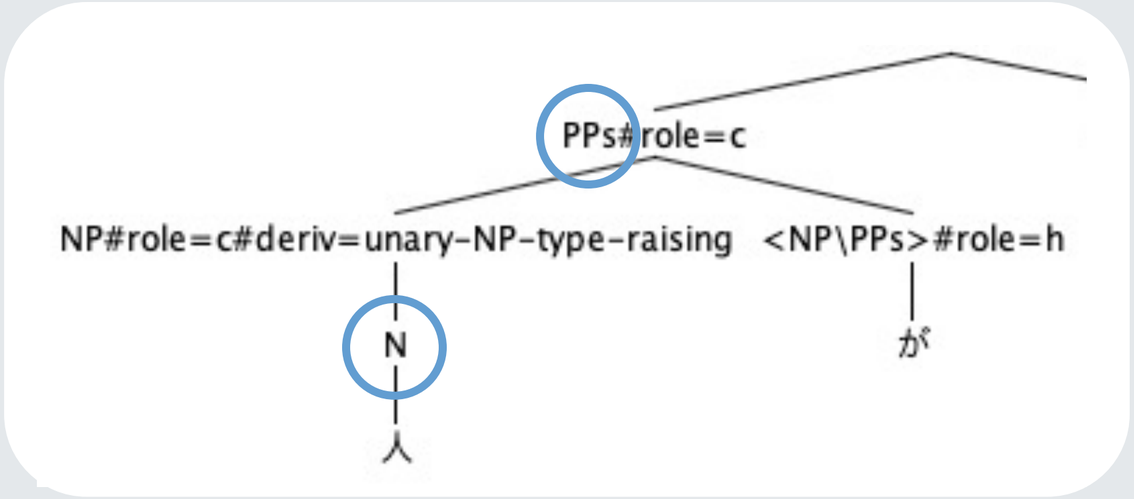




ABCツリーバンク [Kubota+, 2019] 3/5

ABC文法の原子的カテゴリ

統語範疇	名称	素性
CP	節	f, t, t-sbj, q, q-sbj, x
FRAG	断片	
N	**名詞**	**s**
NP	名詞句	q
NPR	固有名詞句	
NUM	数量詞	
NUMCLP	数量詞句	
PP	**後置詞句**	**s, s2, o1, o2**
PRO	代名詞	
Q	量化詞	
S	節（IP相当）	a, e, imp, m, nml, rel, smc, sub
WNUM	疑問数量詞	
WPRO	疑問代名詞	



ABCツリーバンク [Kubota+, 2019] 4/5

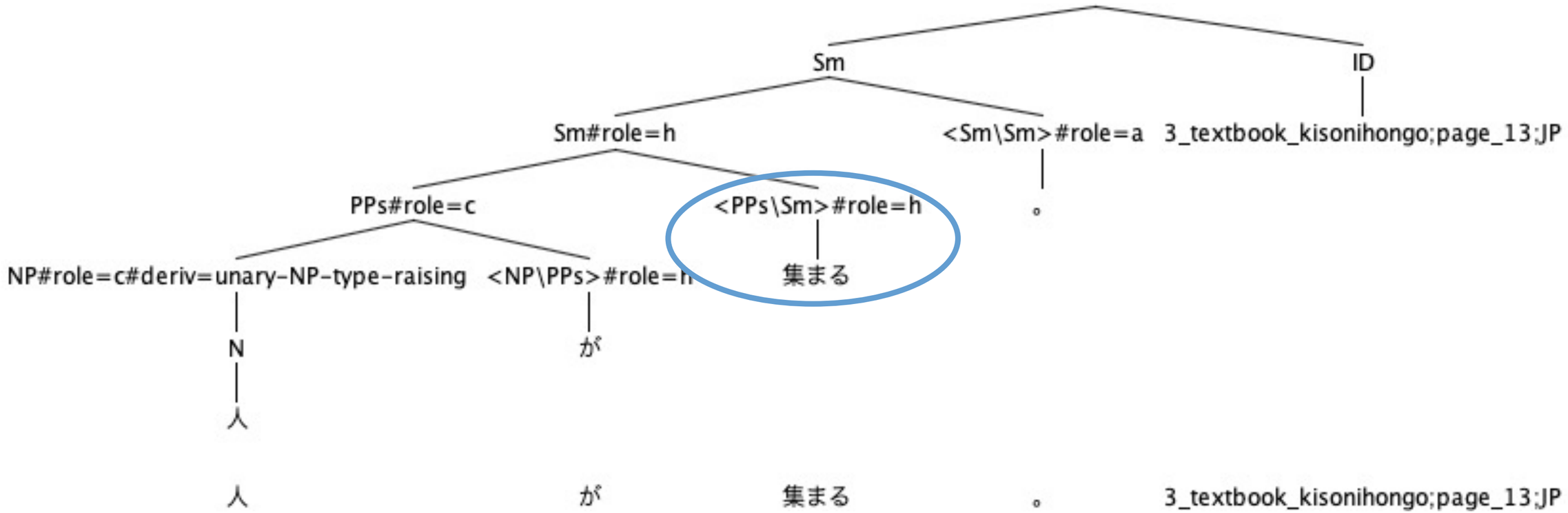
特徴

日本語CCGbankの項構造などの課題点の多くが改善された

課題

活用の種別などの統語情報が不十分

ABCツリーバンク [Kubota+, 2019] 5/5



先行研究

01

日本語CCGbank
[Uematsu+ 2013]

02

ABC**ツリーバンク**
[Kubota+ 2019]

03

lightblue
[Bekki+ 2016]

lightblue [Bekki+, 2016] 1/5

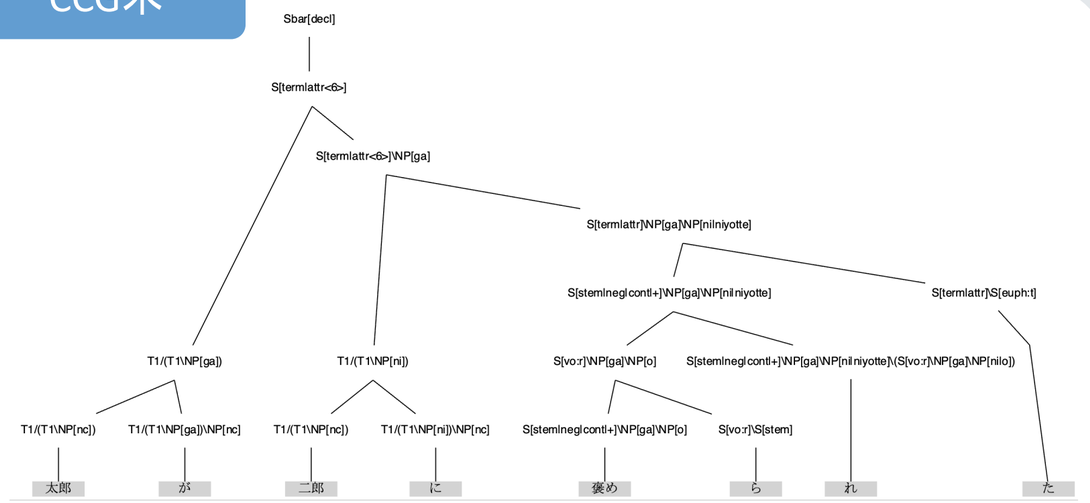
- CCG統語解析器
- 格フレームを元に作られた語彙辞書とCCGの組合せ規則に基づいてCCG統語解析を行う

自然言語

「太郎が二郎に褒められた」

lightblue

CCG木

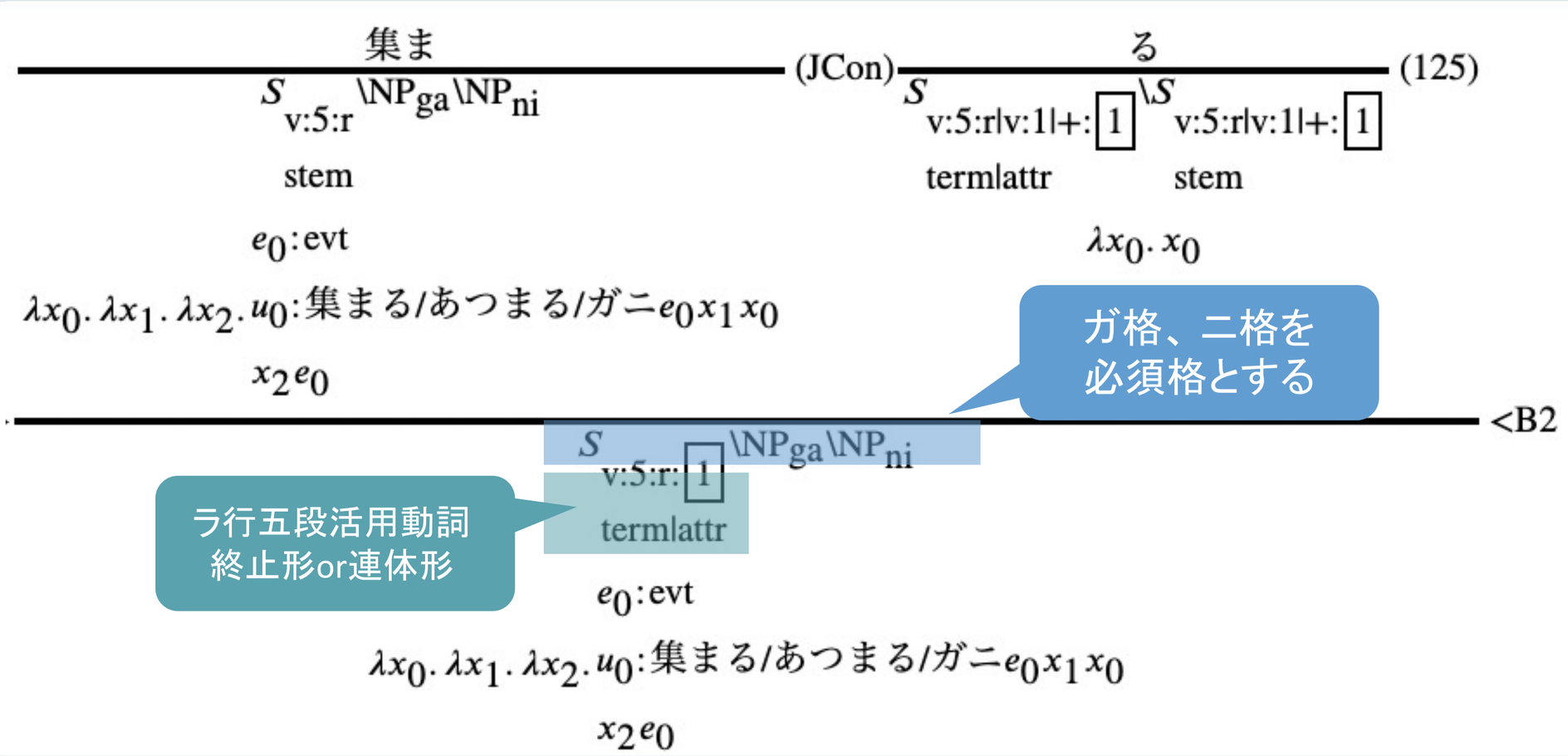


lightblue [Bekki+, 2016] 3/5

特徴

詳細な統語素性を含むCCG木を出力する

lightblue [Bekki+, 2016] 4/5



lightblue [Bekki+, 2016] 5/5

特徴

詳細な統語素性を含むCCG木を出力する

課題

項構造に関する誤りが多い

- 本来自然ではない用言が選択される
- 用言の格フレームに間違いが含まれる

03

提案手法

提案手法

ABCツリーバンクの特徴

項構造が人手によって
正確に記述されている

＋

lightblueの特徴

出力に**詳細な統語素性**を
与えることができる

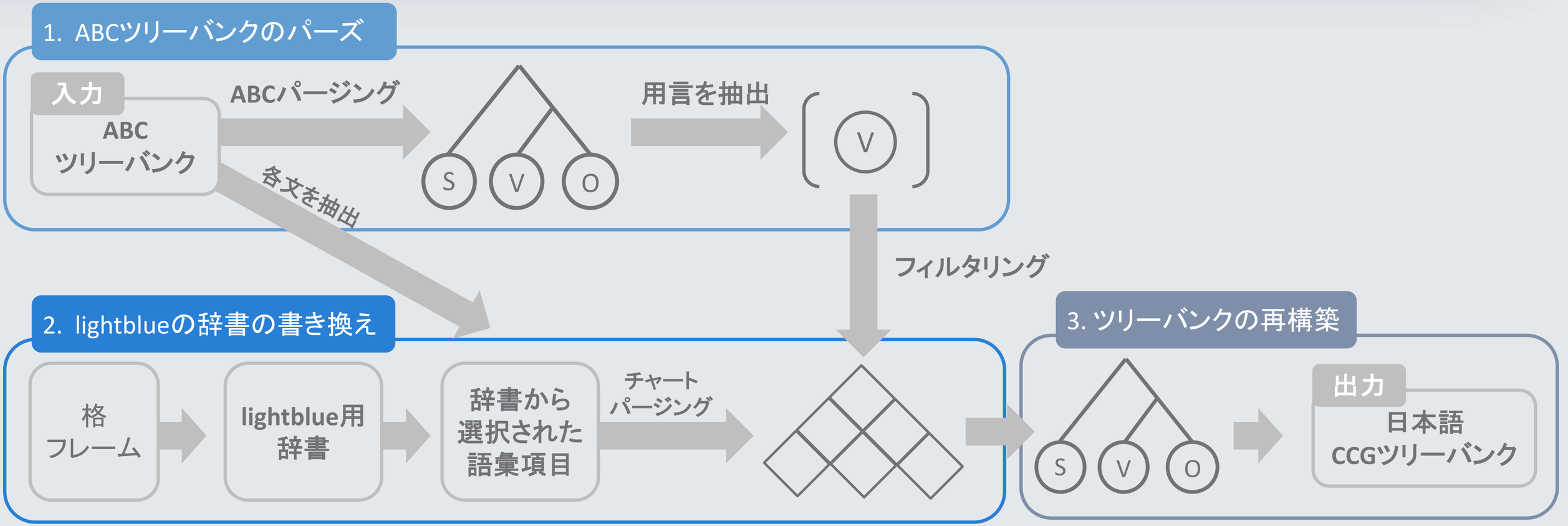
言語学的に妥当で詳細な情報を持った
ツリーバンクを構築する

ABCツリーバンクのリフォージングの流れ

リフォージング：ツリーバンクを分解し再構築する手法

1. ABCツリーバンクのパーズ
2. lightblueの辞書の書き換え
3. ツリーバンクの再構築

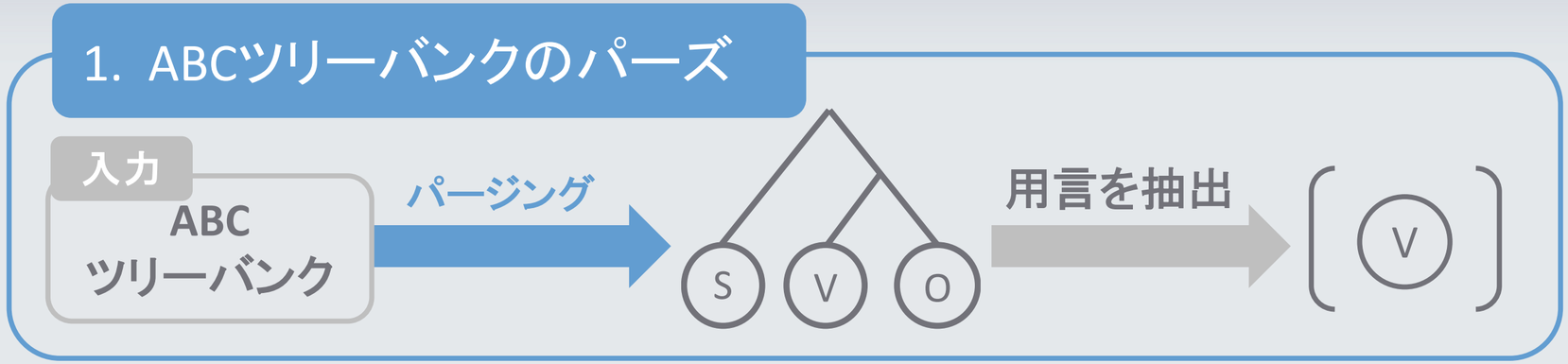
ABCツリーバンクのリフォージングの詳細



リフォージング –ABCツリーバンクのパーズ 1/3

1. ABCツリーバンクのパーズ
2. lightblueの辞書の書き換え
3. ツリーバンクの再構築

リフォージング –ABCツリーバンクのパーズ 2/3



ABCツリーバンクのパーザを作成し木構造データで扱えるようにする

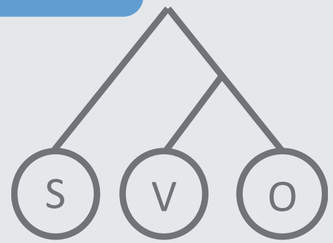
リフォージング –ABCツリーバンクのパーズ 3/3

1. ABCツリーバンクのパーズ

入力

ABC
ツリーバンク

パージング



用言を抽出

例

入力

「会議が始まった」

出力

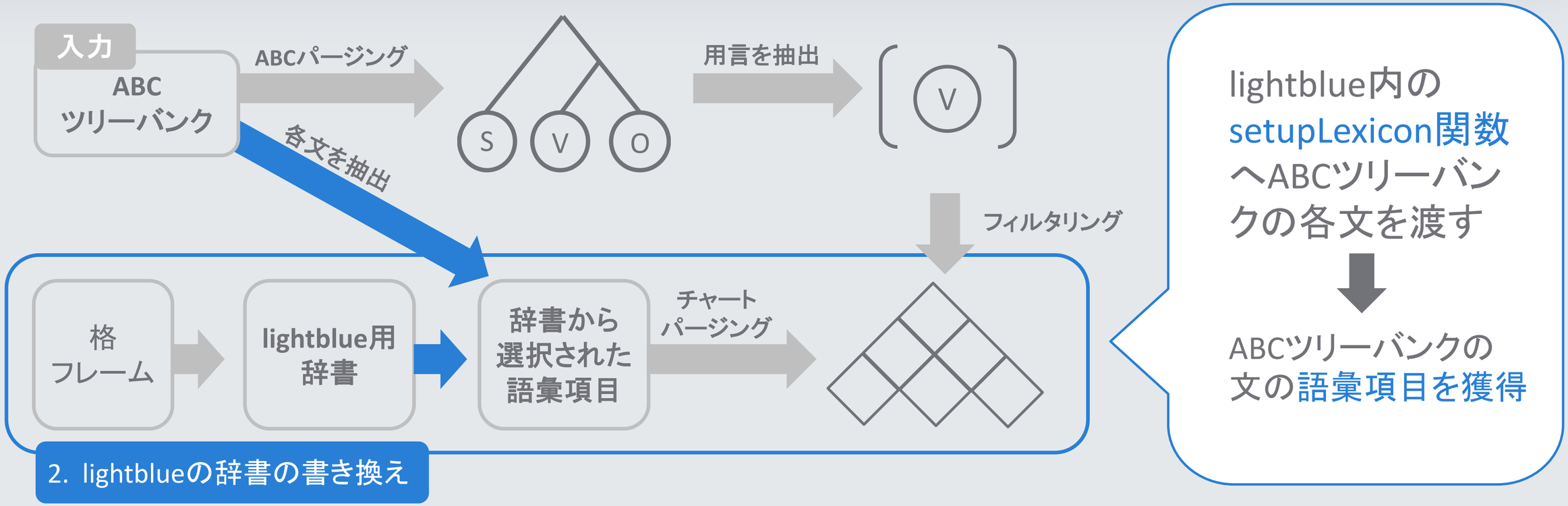
[(“始まっ”, PPs\Sm, 3, 5)]

表層形　統語範疇　開始位置　終了位置

リフォージング – lightblueの辞書の書き換え

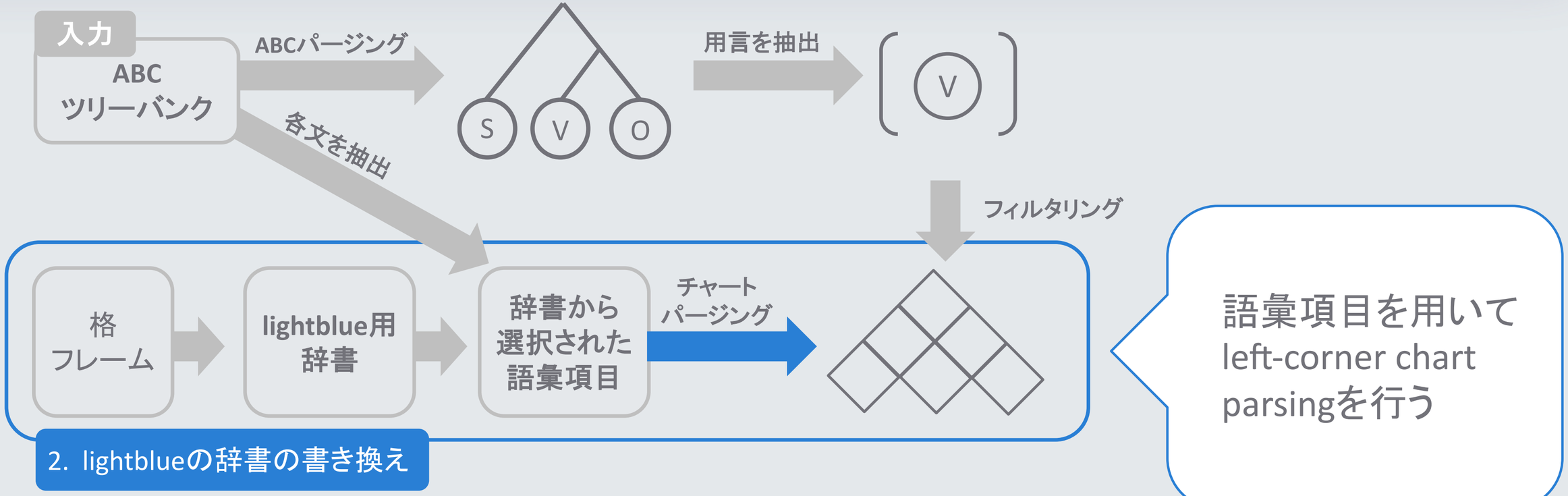
1. ABCツリーバンクのパーズ
2. lightblueの辞書の書き換え
3. ツリーバンクの再構築

リフォージング – lightblueの辞書の書き換え 2/9

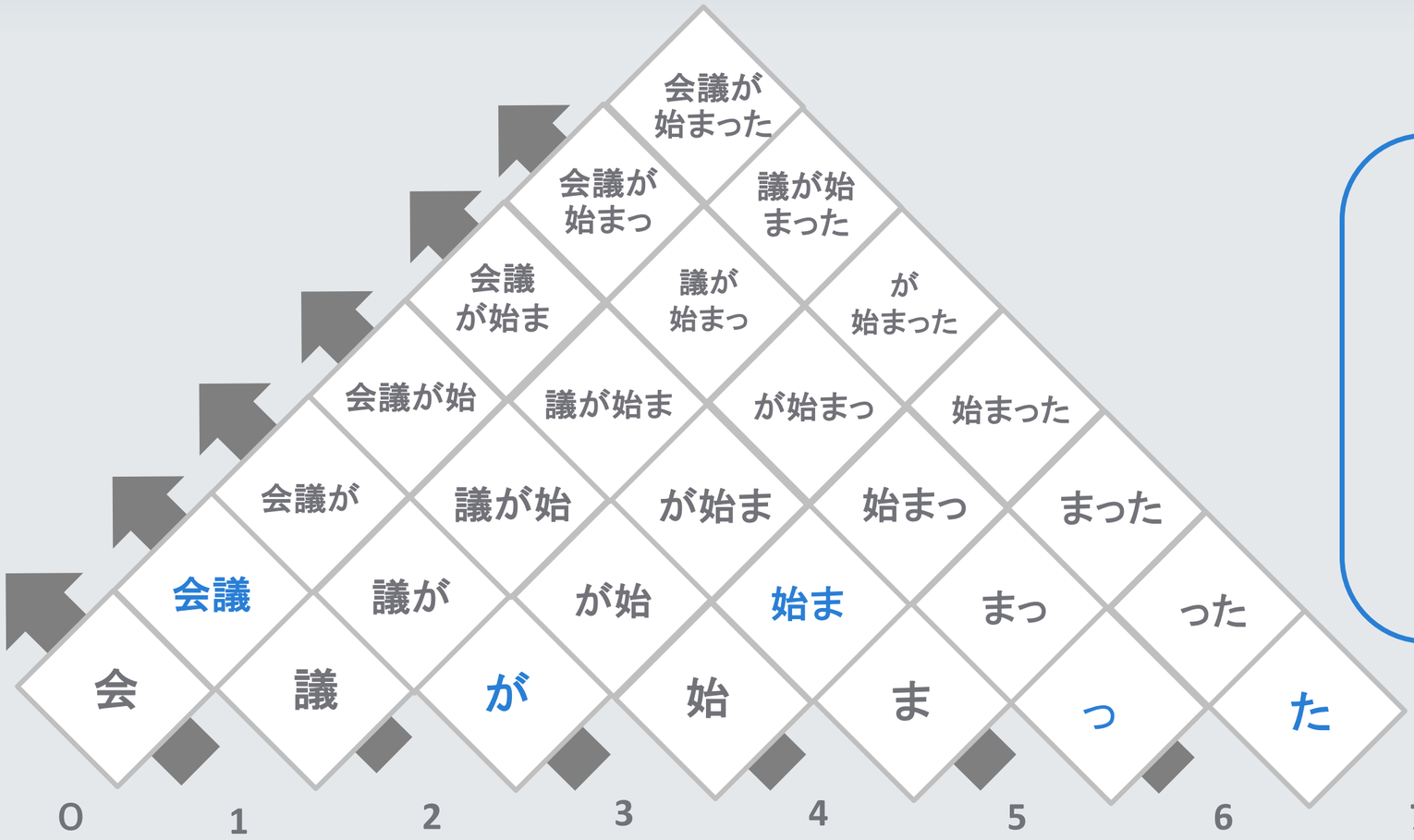


setupLexicon関数 :文を受け取ると、受け取った文を解析するために必要な語彙項目のリストを返す関数

リフォージング – lightblueの辞書の書き換え 3/9



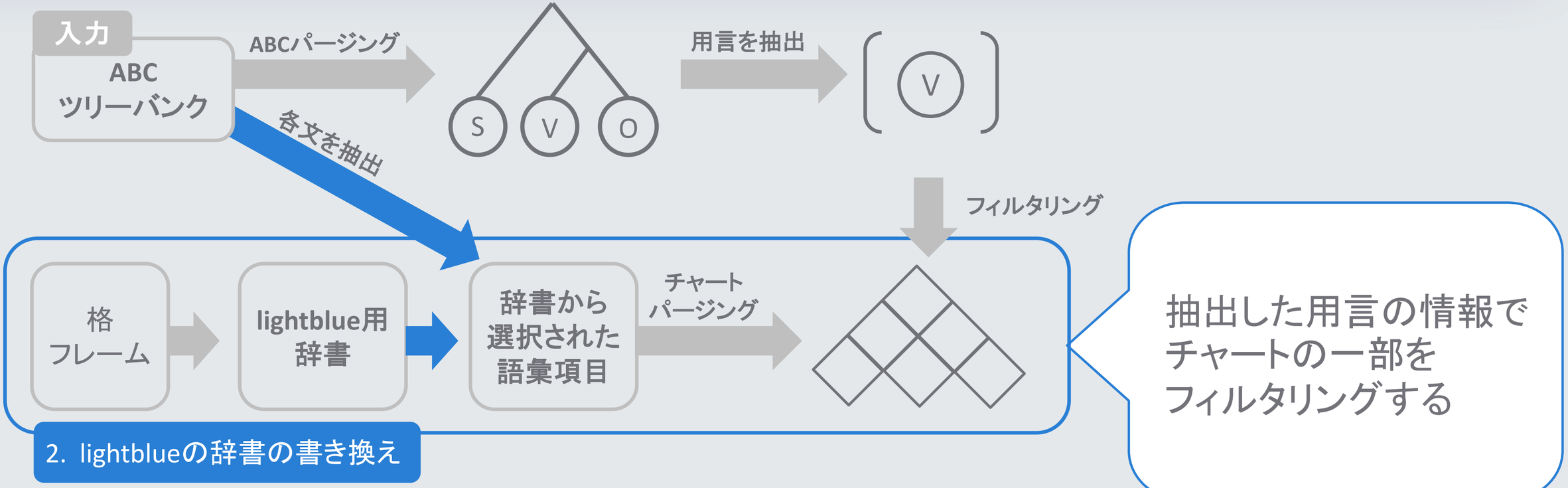
Left-corner chart parsing



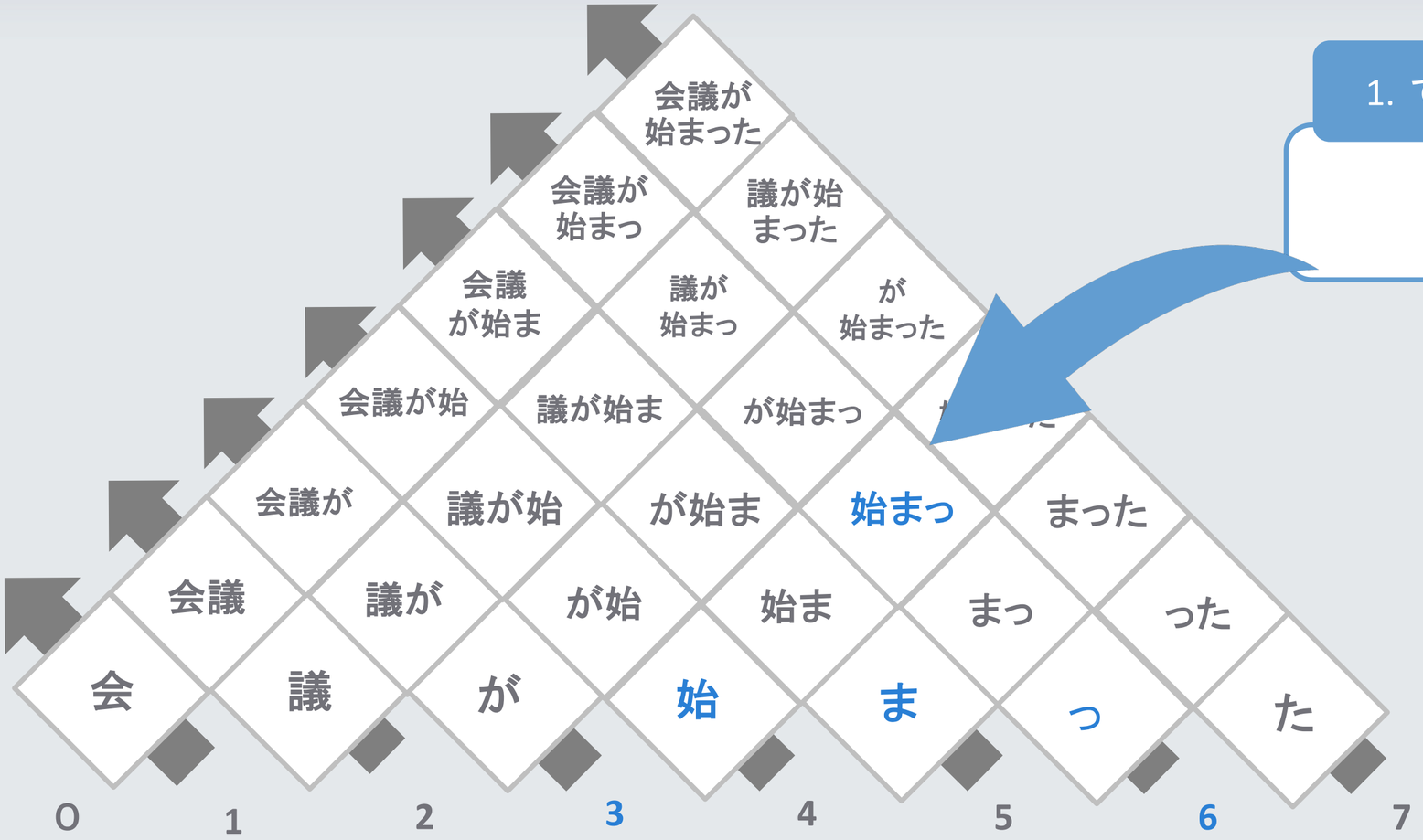
抽出した文の部分文字列
全ての組み合わせにスコアをつける

スコアの高い組み合わせを出力する

リフォージング – lightblueの辞書の書き換え 5/9



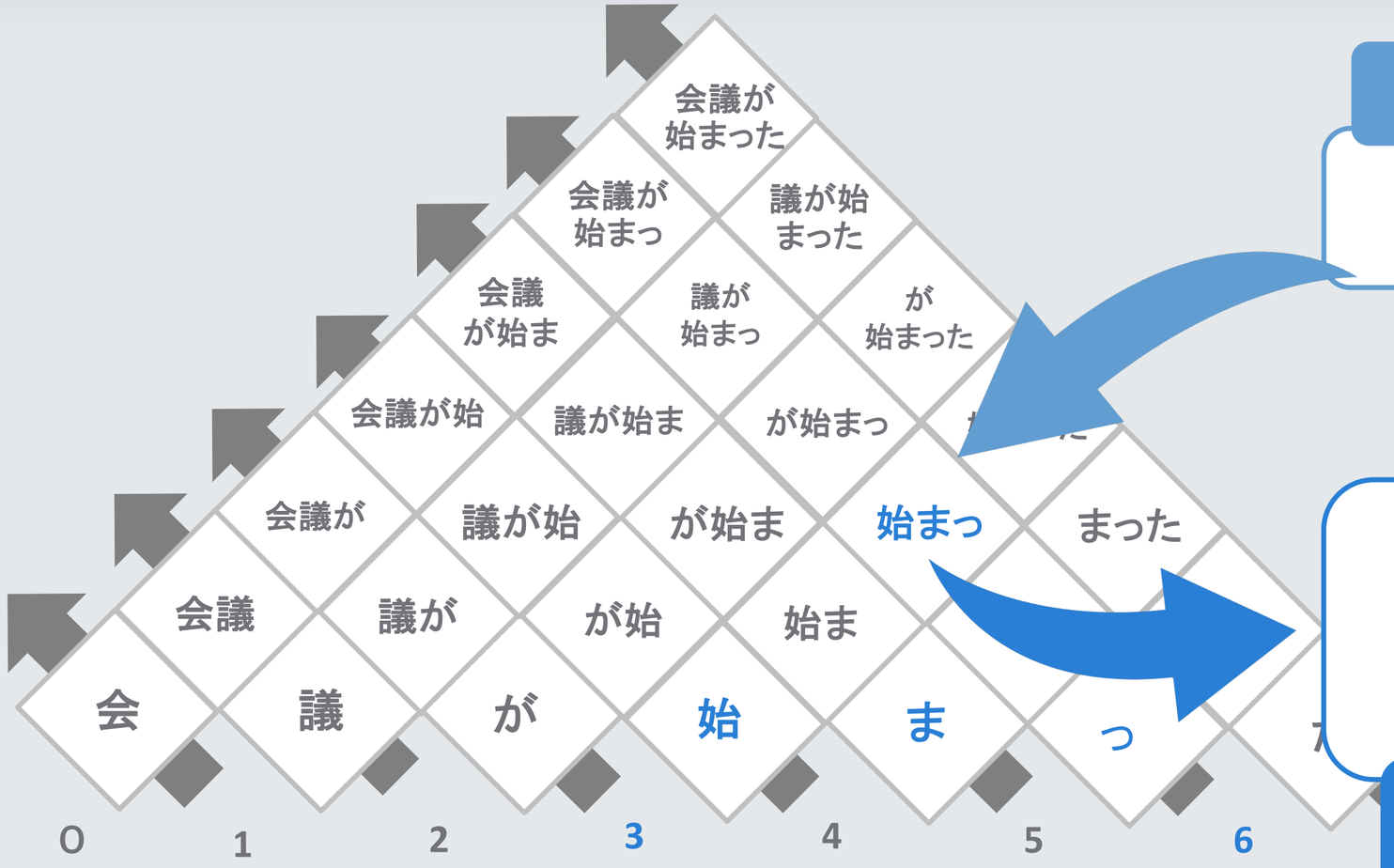
リフォージング – lightblueの辞書の書き換え 6/9



1. で抽出した用言の情報

[("始まっ", PPs\Sm,3,5)]

リフォージング – lightblueの辞書の書き換え 7/9



1. で抽出した用言の情報

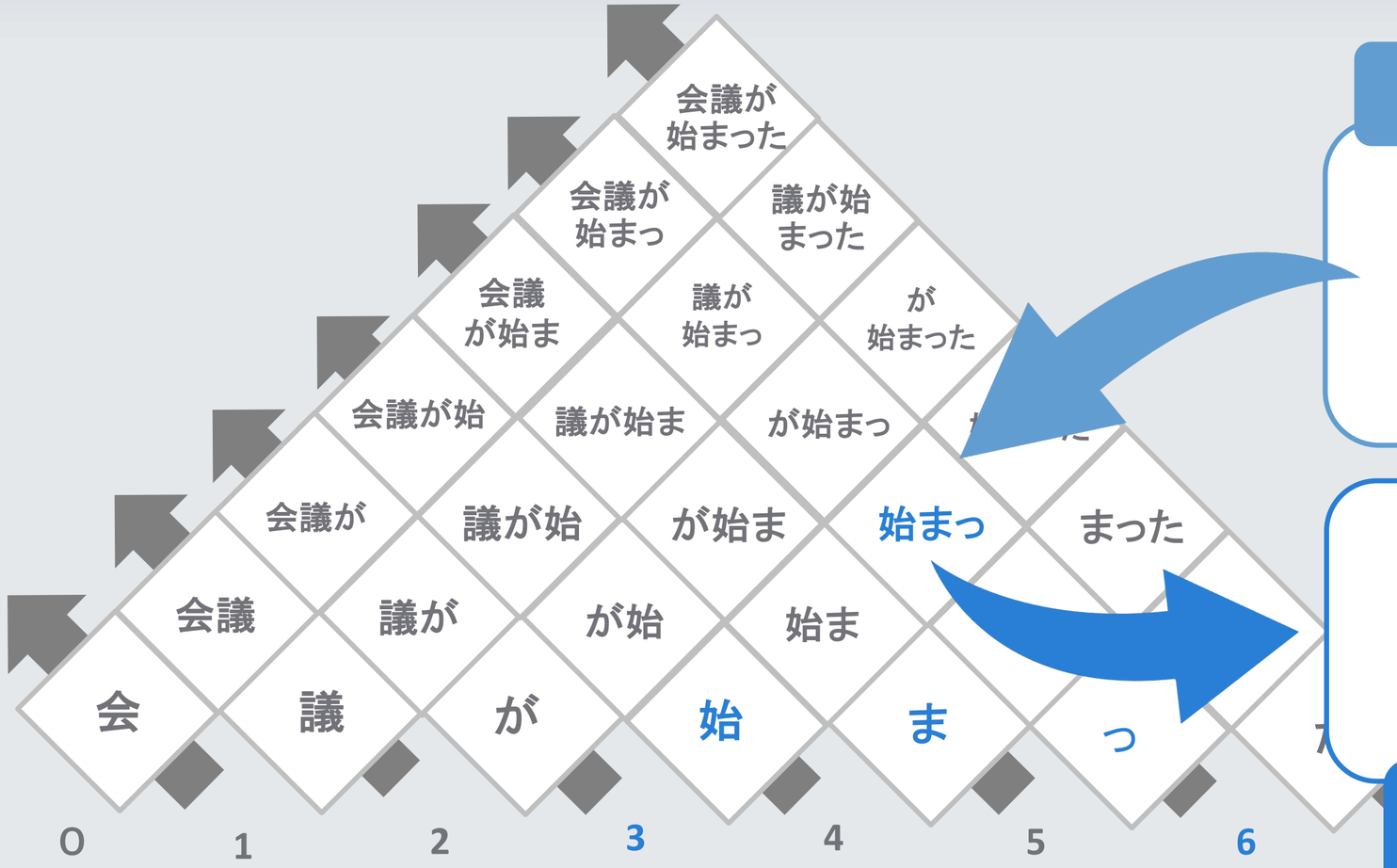
[("始まっ", PPs\Sm,3,5)]

$S_{[v:5:r<1>][euph:t]} \setminus NP_{ga} \setminus NP_{ni}$

$S_{[v:5:r<1>][euph:t]} \setminus NP_{ga}$

lightblueの辞書に登録されている
語彙項目の統語範疇

リフォージング – lightblueの辞書の書き換え 8/9



1. で抽出した用言の情報

[("始まっ", PPs\Sm,3,5)]

lightblueの統語範疇へ変換

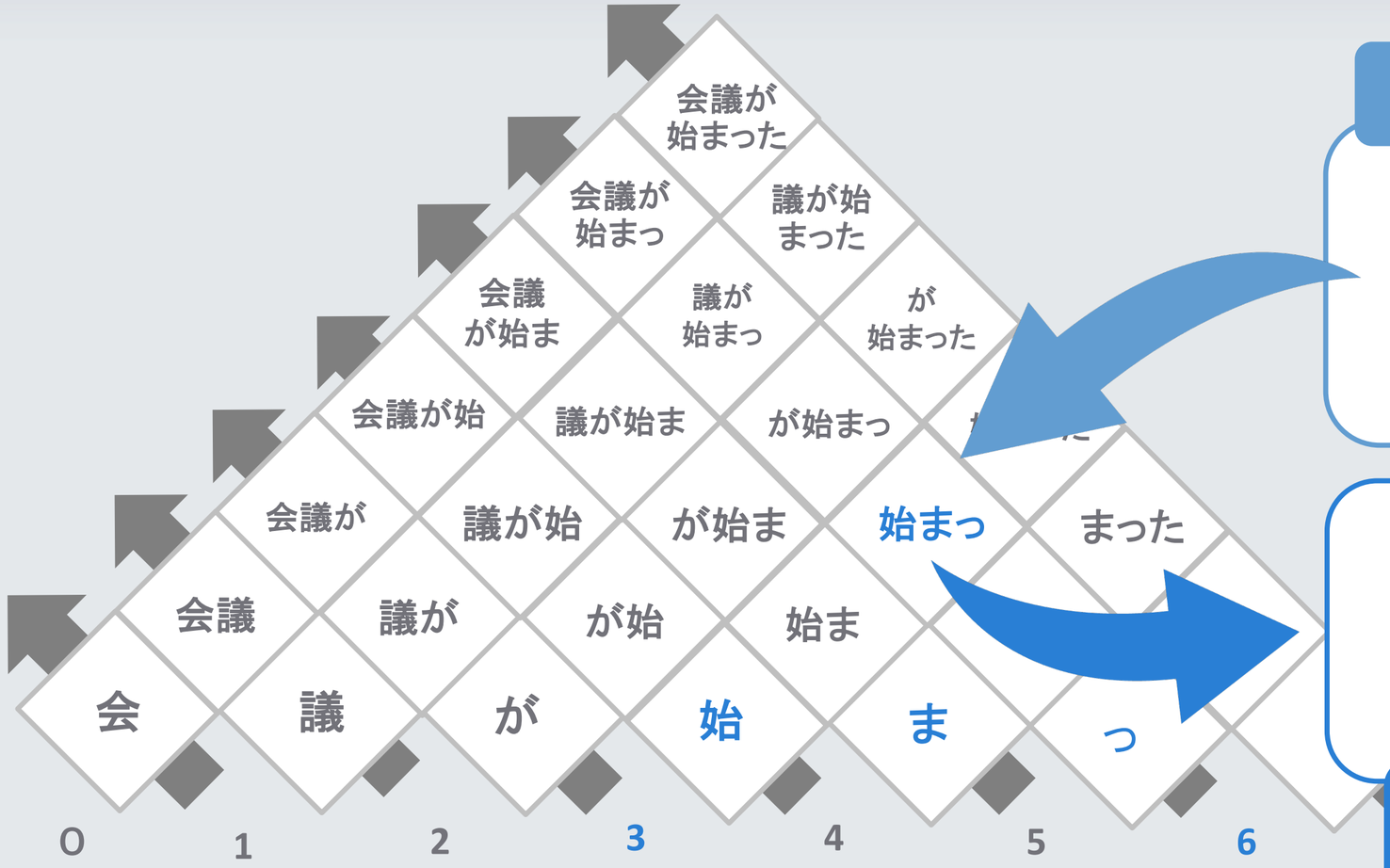
$S \setminus NP_{ga}$

$S_{[v:5:r<1>][euph:t]} \setminus NP_{ga} \setminus NP_{ni}$

$S_{[v:5:r<1>][euph:t]} \setminus NP_{ga}$

lightblueの辞書に登録されている
語彙項目の統語範疇

リフォージング – lightblueの辞書の書き換え 9/9



1. で抽出した用言の情報

[("始まっ", PPs\Sm,3,5)]

lightblueの統語範疇へ変換

$S \setminus NP_{ga}$

~~$S_{[v:5:r<1>][euph:t]} \setminus NP_{ga} \setminus NP_{ni}$~~

$S_{[v:5:r<1>][euph:t]} \setminus NP_{ga}$

lightblueの辞書に登録されている
語彙項目の統語範疇

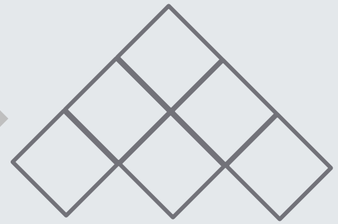
リフォージング –ツリーバンクの再構築 1/3

1. ABCツリーバンクのパーズ
2. lightblueの辞書の書き換え
3. ツリーバンクの再構築

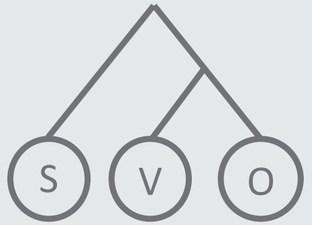
リフォージング –ツリーバンクの再構築 2/3

辞書から
選択された
語彙項目

チャート
パージング



3. ツリーバンクの再構築



出力

日本語
CCGツリーバンク

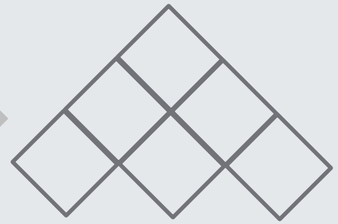
フィルターされたチャートを用いてABCツリーバンクの文の統語解析を行う

成功→ CCG統語構造が出力
失敗→ エラーが出力

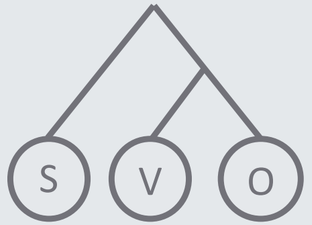
リフォージング –ツリーバンクの再構築 3/3

辞書から
選択された
語彙項目

チャート
パージング



3. ツリーバンクの再構築



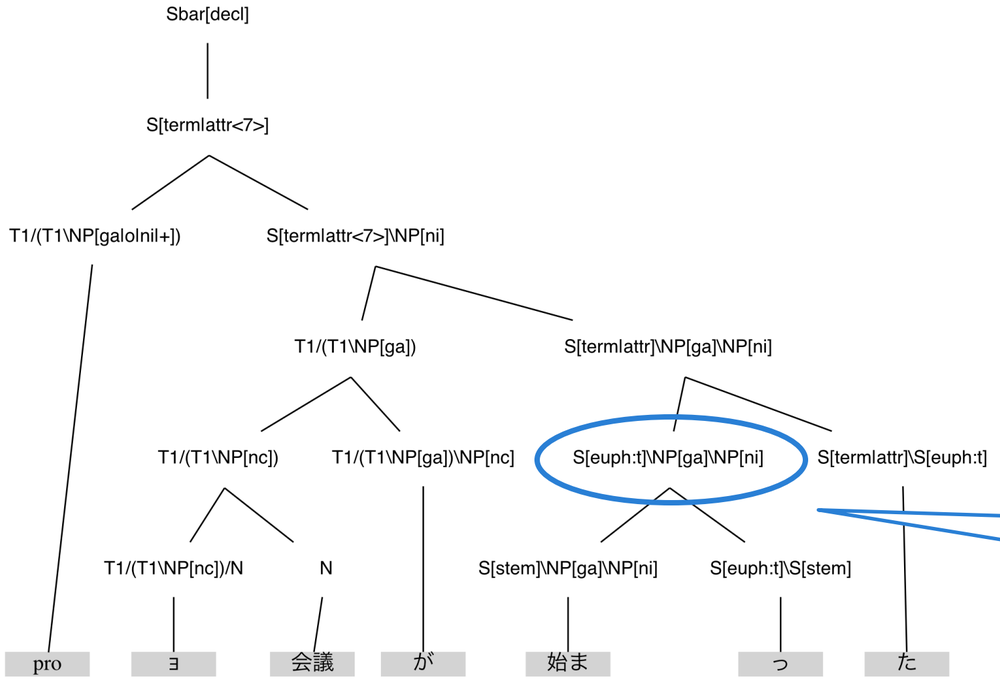
出力

日本語
CCGツリーバンク

ABCツリーバンクの**全ての文**で統語解析を行い
ツリーバンクとしてまとめる

リフォージングの成功例

フィルタリング前



lightblueの辞書に登録されている語彙項目の統語範疇

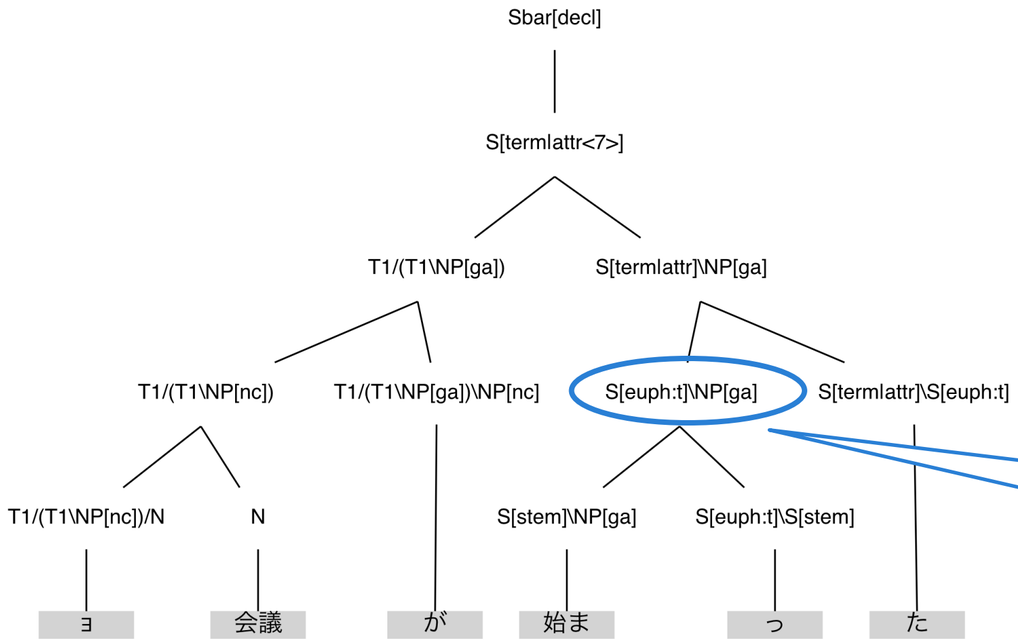
$$S_{[v:5:r<1>][euph:t]} \setminus NP_{ga} \setminus NP_{ni}$$

$$S_{[v:5:r<1>][euph:t]} \setminus NP_{ga}$$

必須格はガ格のみだが、ニ格も必須格としている

リフォージングの成功例

フィルタリング後



lightblueの辞書に登録されている語彙項目の統語範疇

~~$S_{[v:5:r<1>][euph:t]} \setminus NP_{ga} \setminus NP_{ni}$~~

$S_{[v:5:r<1>][euph:t]} \setminus NP_{ga}$

必須格がガ格になっている

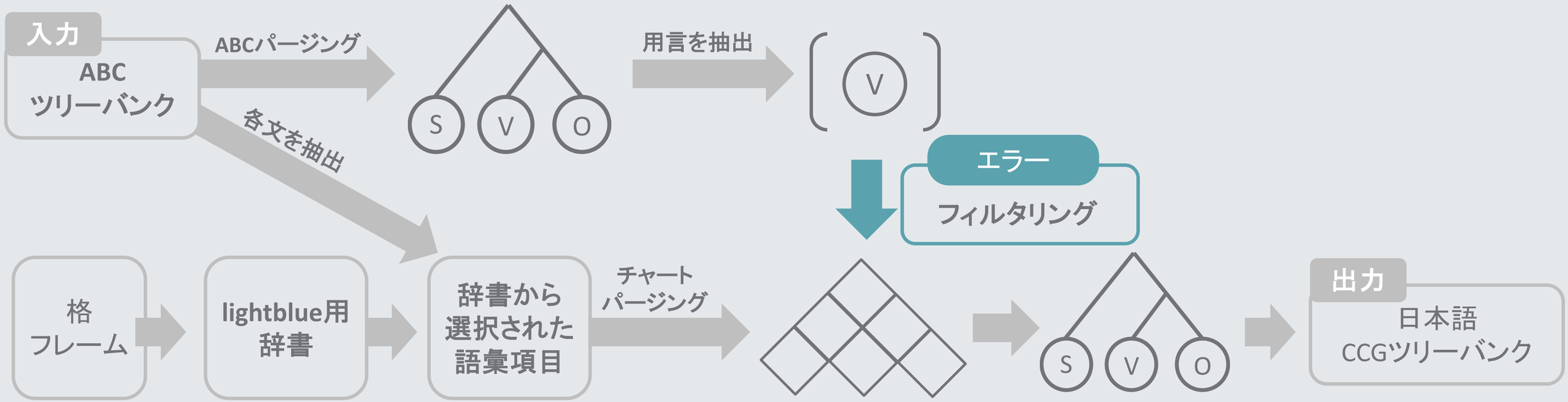
04

エラー分析

エラー分析

課題

lightblueの用言の語彙項目を 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リングする際エラーが発生



エラー分析

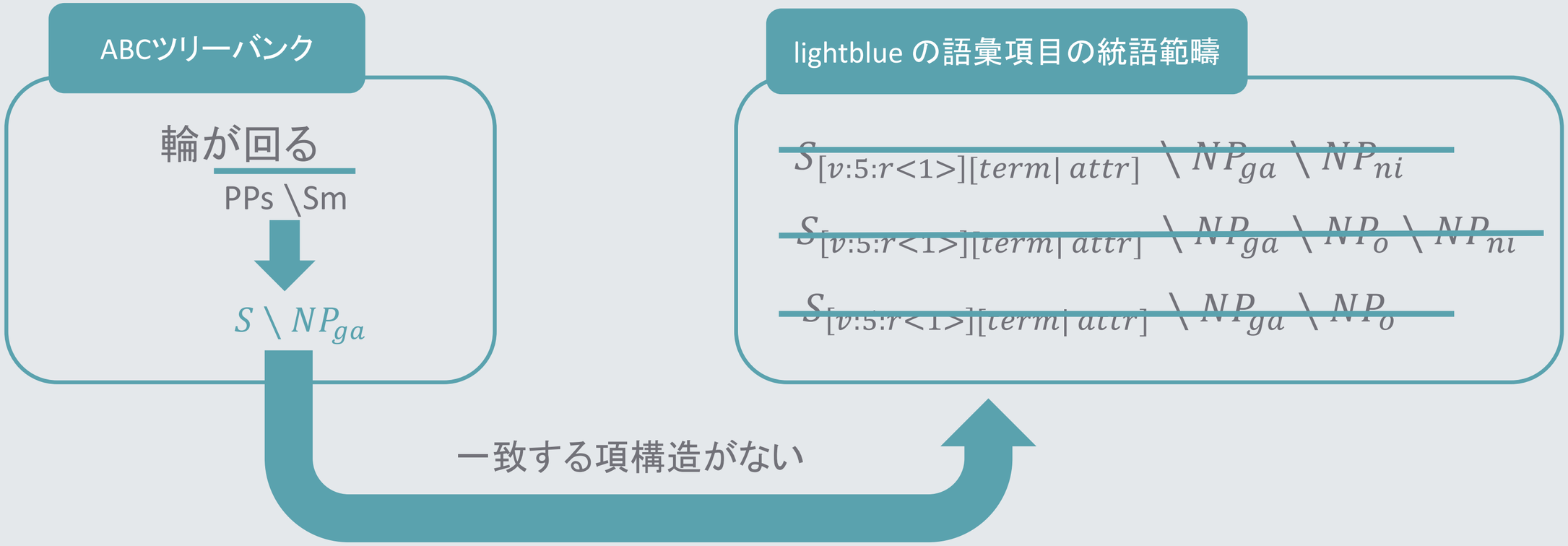
課題

lightblueの用言の語彙項目を 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リングする際エラーが発生

01 lightblue の語彙項目に正しい用言が含まれない ケース

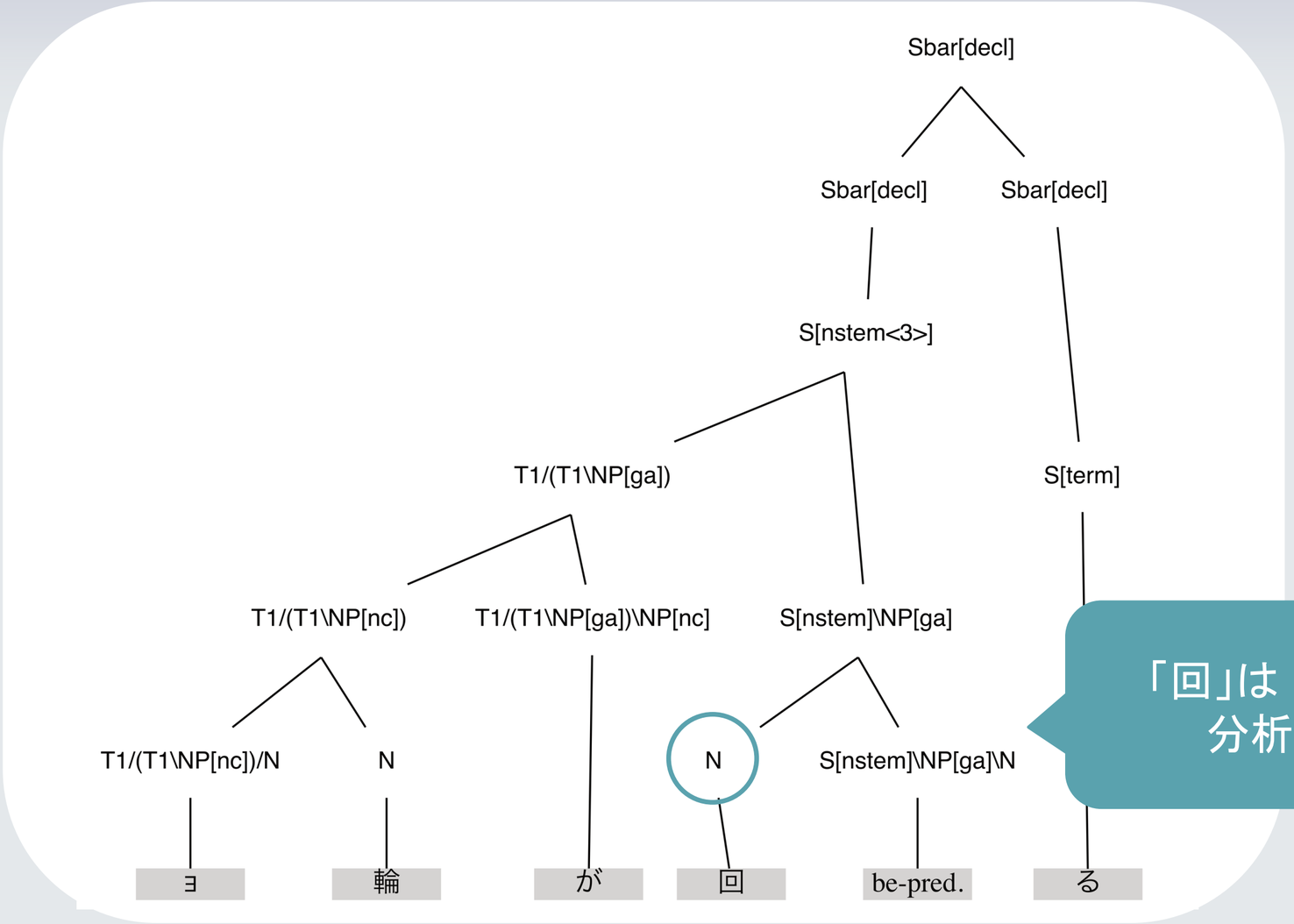
02 ABC ツリーバンクの項構造が誤っているケース

エラー分析 - lightblueの語彙項目に正しい用言が含まれないケース



「回る」の語彙項目が全て削除されてしまう

エラー分析 ABCツリーバンクの項構造が誤っているケース



「回」は 名詞 として
分析されている

エラー分析 – ABCツリーバンクの項構造が誤っているケース

ABCツリーバンク

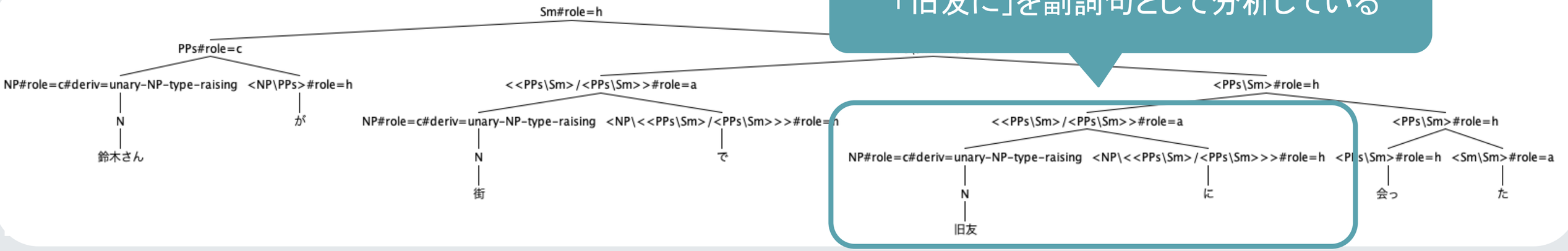
鈴木さんが街で旧友に会った。

PPs \ Sm

$S \setminus NP_{ga}$

「会う」の必須格はガ格、ニ格になるべきである

「旧友に」を副詞句として分析している



エラー分析 – ABCツリーバンクの項構造が誤っているケース

ABCツリーバンク

鈴木さんが街で旧友に会った。

PPs \ Sm

$S \setminus NP_{ga}$

lightblue の語彙項目の項構造

~~$S_{[v:5:w<1>][euph:t]} \setminus NP_{ga} \setminus NP_{ni} \setminus Sbar_{ga}$~~

~~$S_{[v:5:w<1>][euph:t]} \setminus NP_{ga} \setminus Sbar_{ga}$~~

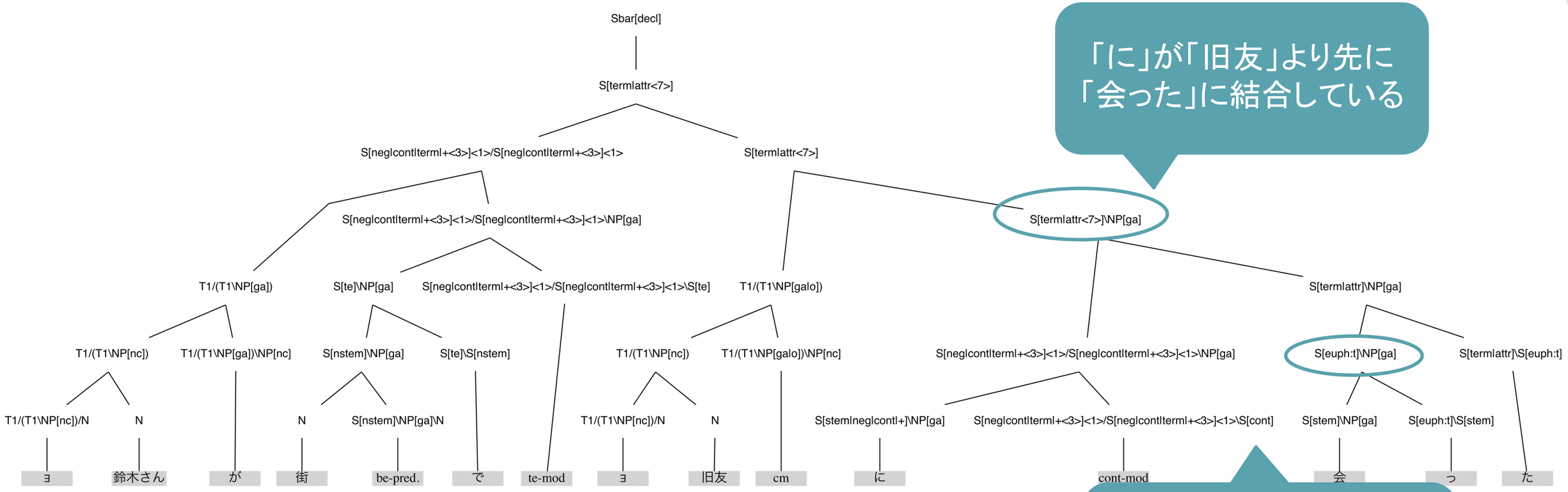
~~$S_{[v:5:w<1>][euph:t]} \setminus NP_{ga} \setminus NP_{ni}$~~

$S_{[v:5:w<1>][euph:t]} \setminus NP_{ga}$

本来は $S \setminus NP_{ga} \setminus NP_{ni}$ が残るべき

正しい語彙項目が削除されてしまっている

エラー分析 ABCツリーバンクの項構造が誤っているケース



「に」が「旧友」より先に
「会った」に結合している

「会っ」がガ格のみ
取るようになっている

05

まとめ

まとめ

01 **ABCツリーバンク** と **lightblue** が持つ利点を組み合わせることで
言語学的に妥当で詳細な統語情報を持った**日本語CCGツリーバンクを構築する手法(リフォージング)**を提案した。

02 リフォージングによって**正しい日本語CCG統語構造が部分的に得られた**が、誤りが含まれるケースも残されている

今後の展望

より正しい CCG 統語構造を出力できるようにフィルタリングのアルゴリズムを改良する

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 (to appear), the workshop in the Georgetown University Round Table on Linguistics 2023 (GURT2023), forthcoming.
3. Daisuke Kawahara and Sadao Kurohashi. A fullylexicalized probabilistic model for Japanese syntactic and case structure analysis. In Proc. of the Human Language Technology Conference of the NAACL, Main Conference, June 2006.
4. 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.
5. Mark Steedman. The Syntactic Process. MIT Press, 2000.
6. Mark J. Steedman. Surface Structure and Interpretation. The MIT Press, Cambridge, 1996.
7. 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.
8. 戸次大介. 日本語文法の形式理論. くろしお出版, 東京, 2010.
9. 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.
10. 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.
11. 花岡洋輝, 増田勝也, 植松すみれ, 美馬秀樹. 日本語助詞「と」コーパスの構築.
    言語処理学会 第 18 回年次大会 発 表論文集 (2012 年 3 月), pp. 247–250. 一般社団法人 言 語処理学会, 2012.

補足 – 語彙項目、統語素生

語彙項目

辞書に記述される語と統語範疇の対応付け

例: Keats ⊢ NP

eats ⊢ $(S \setminus NP) / NP$

apples ⊢ NP

統語素性

統語範疇についてのより詳しい指定

例: eats ⊢ $(S \setminus NP_{nom}) / NP_{acc}$

補足 – 格フレーム

格フレーム

私がカメラで走っている少女を見た。

計算機には

- カメラで見た
- カメラで走っている

のどちらが妥当かを判定するのが難しい

補足 – 格フレーム

格フレーム

用言とそれに関係する名詞を用言の各用法ごとに整理したもの

例:「とまる」

1.「停止する」の意味の「とまる」

{ 車, バス, タクシー, ...} が { 駐車場, 横, 路肩, ...} に　とまる

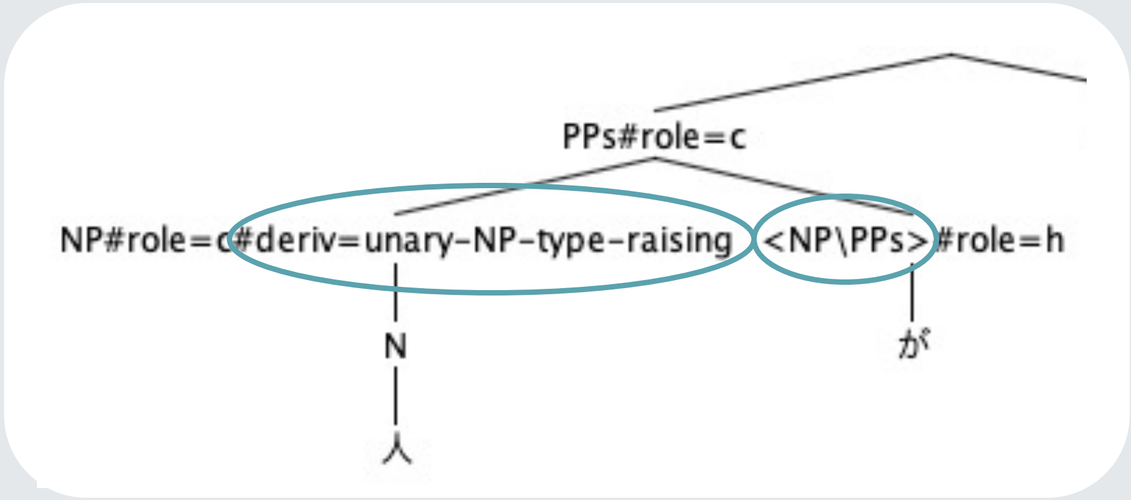
2.「吊るす」の意味の「かける」

{ 人, 3人, ...} が { ホテル, 家, ...}に {3人, 家族, ...} で とまる

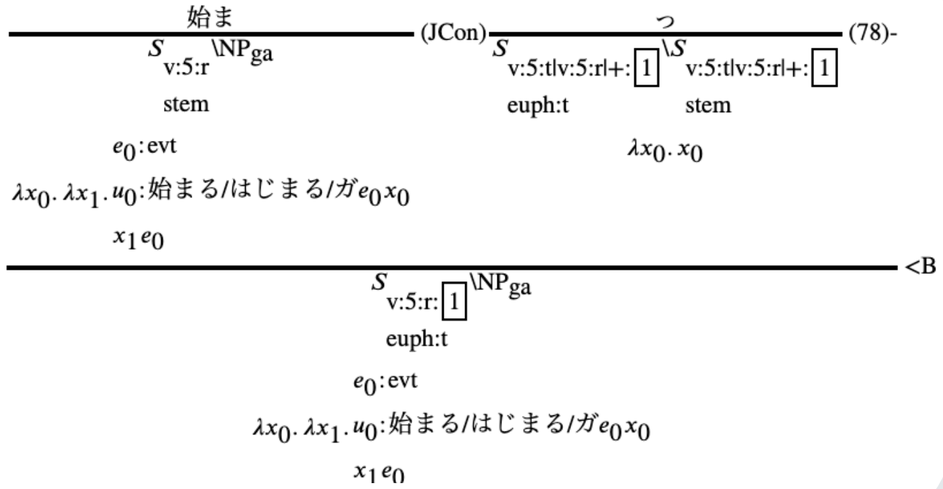
補足 – ABC文法の代表的な規則

ABC文法の代表的な規則

名称	規則	
関数適用（順方向）	A/B B ⇒ A	ABC文法
関数適用（逆方向）	B B\A ⇒ A	
関数合成（順方向）	A/B (B/C)/$ ⇒ (A/C)/$	
関数合成（逆方向）	\ (C\B)C\A⇒\ (C\A)	
名詞の名詞句化	N ⇒ NP	unary規則
名詞句の述語化	NP ⇒ PPs\S	
名詞句の後置詞句化	NP ⇒ PPs; NP ⇒ PPo1; etc.	
関係節による名詞修飾	PPs\Srel ⇒ N/N; PPo1\Srel ⇒ N/N; Srel ⇒ N/N; etc.	
連用節化	Sa ⇒ S/S	
名詞の副詞化	NP ⇒ S/S; NP ⇒ (PPs\S)/(PPs\S); etc.	
スクランブリング	PPo1\ (PPs\S) ⇒ PPs\ (PPo1\S); etc.	
疑問節化・感嘆節化	Ssub ⇒ CPq; Ssub ⇒ CPx	



補足 – lightblueの辞書の形式



lightblueの辞書の形式

項目	用言「始まっ」の辞書	備考
規則名	BFC1	関数合成（逆方向）
表層形	始まっ	
CCG統語範疇	S[v:5:r<1>][euph:t][\NP[ga]	ラ行五段活用動詞、必須格ガ格、
DTS意味表示	λx0.λx1.(e0:evt) × (u0:始まる/はじまる/ガ(e0,x0)) × x1(e0)	
シグネチャー	[("始まる/はじまる/ガ",(x0:entity)→ (e0:evt)→ type)]	
子ノード	[LEX 始ま S[v:5:r][stem][\NP[ga] λx0.λx1.(e0:evt) × (u0:始まる/はじまる/ガ (e0,x0)) × x1(e0) (JCon) [0.99] Sig. [始まる/はじまる/ガ:(x0:entity)→ (e0:evt)→ type] ,LEX っ S[v:5:t|v:5:r|v:5:w|+<1>][euph:t][] \S[v:5:t|v:5:r|v:5:w|+<1>][stem][] λx0.x0 (78)- [1.00] Sig. []	「始ま」と「っ」の辞書
スコア	99 / 100	0.99
ソース	""	