発表番号 D12-5

CCGによる日本語文処理のモデリング

梶川康平　吉田遼　大関洋平（東京大学）

2023年3月16日 言語処理学会

概要

- 人間は逐次的にどのようにして統語/意味構造を構築しているのか？
- 英語において、

  ① CCGの右枝分かれ構造より左枝分かれ構造の方が、

  ② さらにreveal操作という文処理方略が

  認知的に妥当と主張されている

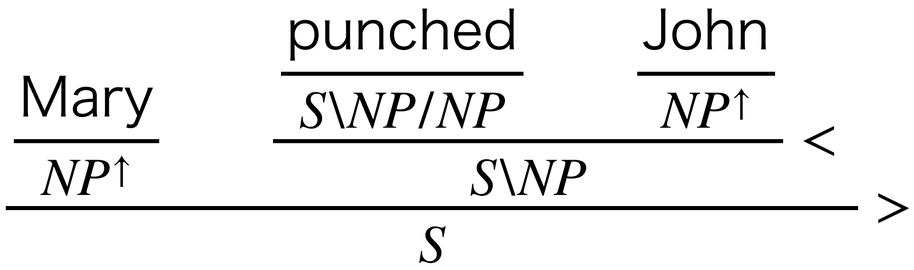
- 言語の構造が異なる日本語においても同様の主張は成り立つのか？
  - 特に、日本語においては②は成り立たないように思われる。

  ➡ ① は成り立ったが、② は成り立たなかった。

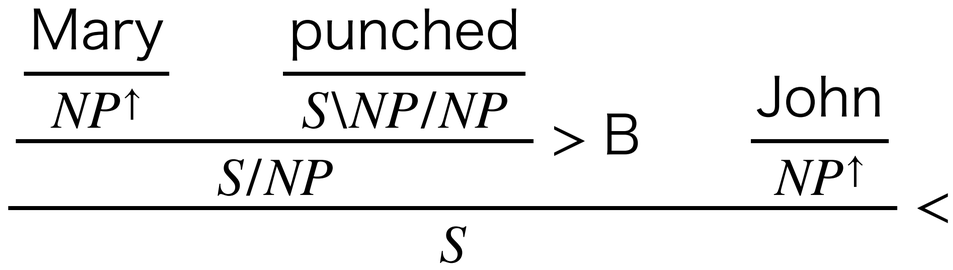
Combinatory Categorial Grammar

Steedman (2000), 戸次 (2010)

- 形式文法として高い記述力をもつ (Joshi, 1985; Stabler, 2013)
- 型依存 (type-dependent) な組合せ規則により統語/意味構造を並行して計算
  - 単語の逐次的な合成による文の導出が可能



右枝分かれ構造

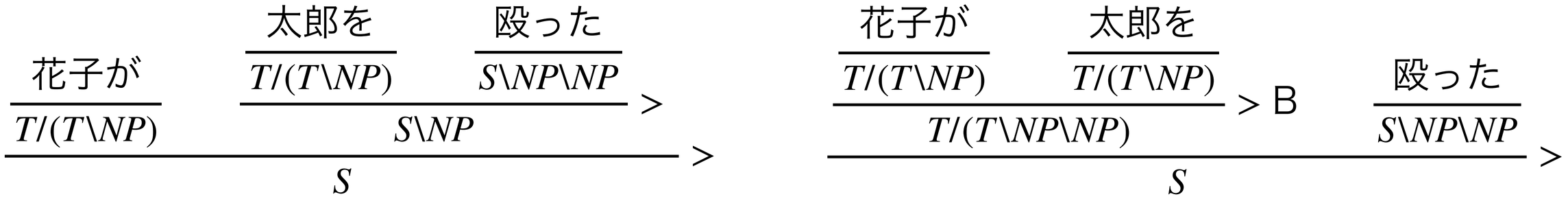


左枝分かれ構造

- 英語における、左枝分かれ構造の(右枝分かれ構造に対する)妥当性は、fMRIデータを通して示されている (Stanojevic' et al., 2021)

日本語の左枝分かれ構造

動詞に先んじた項同士の関係の計算を行う



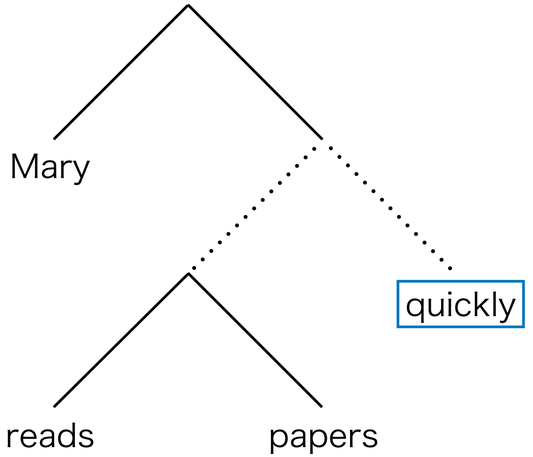
- 心理言語学において、「verb-finalな日本語では、動詞を待たずに項構造が計算されている」と主張されている (Kamide and Mitchell, 1999; Miyamoto, 2002; Isono and Hirose, 2022)
  - 特定の構造に対して検証済み

➡ナチュラリスティックコーパスを使った計算心理言語学的検証

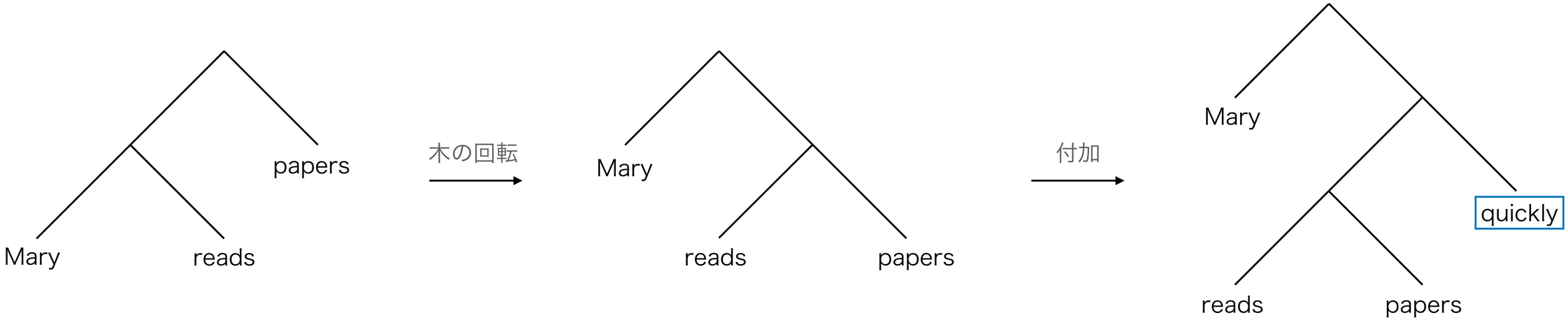
Reveal操作

Stanojevic’ and Steedman (2019)

- 後置修飾詞は、逐次的な文処理の障害となる (e.g., Hale, 2014)

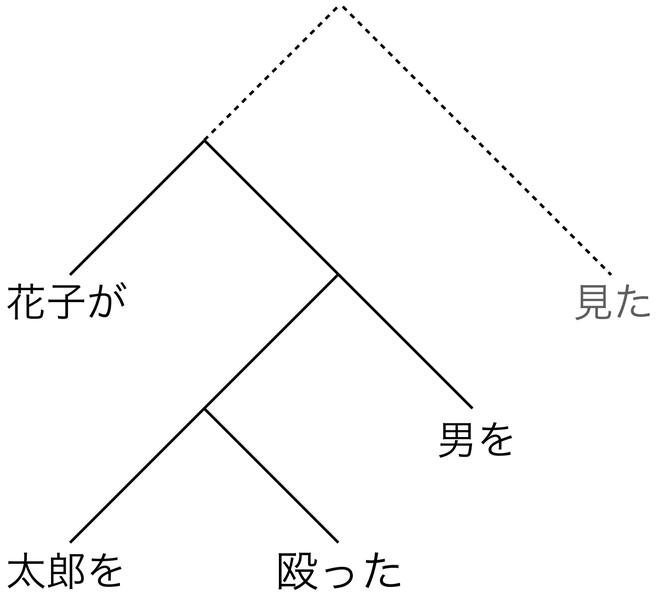


1. 左枝分かれ構造は右枝分かれ構造に変換し (木の回転)
2. 後置修飾詞は既に作った構成素に付加する



- 英語におけるreveal操作の認知的妥当性は、fMRIデータを通して示されている (Stanojevic’ et al., 2021; 2022)

日本語におけるreveal操作



- Reveal操作は日本語においても妥当な操作か？

1. 後置修飾詞がない (Greenberg, 1963)

2. 関係節が埋め込まれた文には、reveal操作を応用できる
    - そのような文は、**典型的なgarden-path文** (井上, 1990)

➡ 日本語の文処理では、reveal操作による予測に反するかも？

実験

方法論

- 検証する仮説:

  日本語において、逐次的に構築している構造として

  ① CCGの右枝分かれ構造より、左枝分かれ構造 が妥当

  ② 左枝分かれ構造より、reveal操作による構造 が妥当

- データ：BCCWJ-EyeTrack (浅原ら, 2019)

  視線走査法による視線情報がアノテーションされたコーパス

- 橋渡し仮説:

  CompositionCount

▸ 尤度比検定で評価

実験

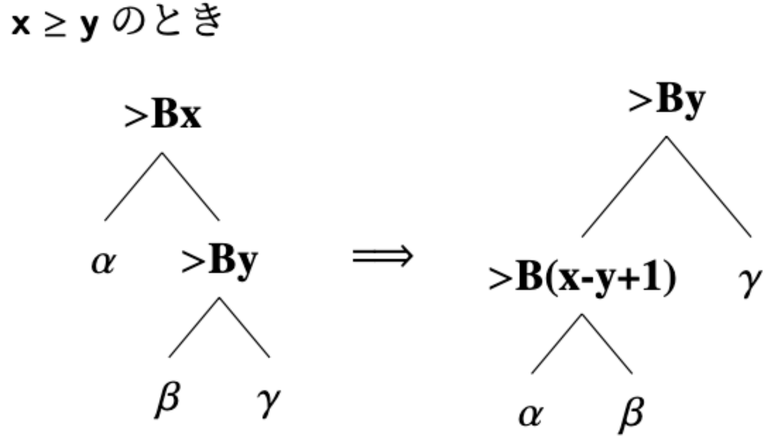
木構造の獲得

右枝分かれ構造：depccg (Yoshikawa et al., 2017) のbest parse

➡ 一部のcategoryを型繰り上げ（>T）

➡ 木を左方向に回転

➡ 左枝分かれ構造



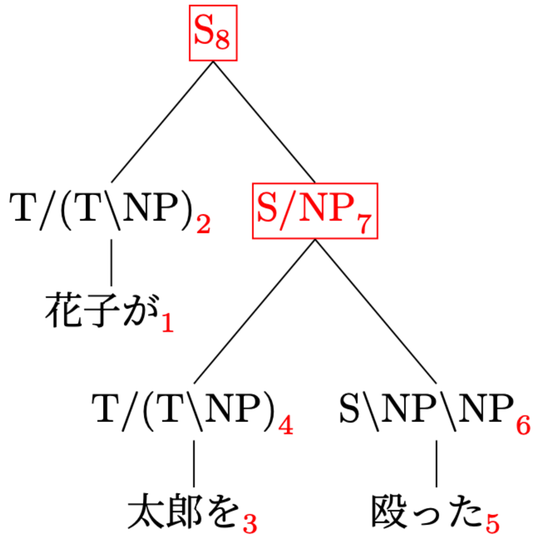
木の回転の例

詳細は予稿集、appendixへ

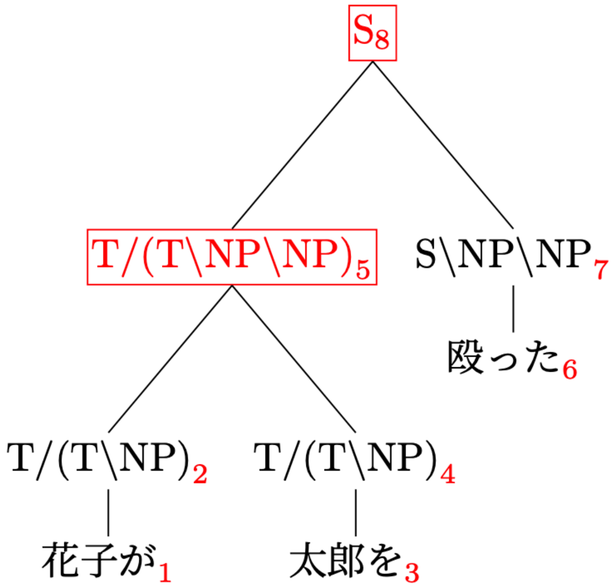
実験

橋渡し仮説：CompositionCount (CC)

- 各単語ごとに新たに構築される二分木の数 = 意味合成の数
  - 先行研究でのNodeCountと異なり、パーサー特有のunary ruleの影響を排した上で、意味計算という側面に注目した指標



CCright: (0, 0, 2)



CCleft: (0, 1, 1)

実験

統計分析

① CompositionCountの有効性検証

② Nested model comparisonによる仮説検証

Baseline:

```
RT ~ dependent + length + frequency + is_first + is_last +
     is_second_last + screenN + lineN + segmentN + (1|article)
     + (1|subj)
```

Data:

視線走査法による総注視時間、13,232/19,176のデータポイントを使用

有意水準:

$\alpha = 0.05/t$（tは検定数）

結果・考察

1. CompositionCountの有効性

	χ^2	df	p
Baseline $<$ All	127.51	3	$<$ **0.0001**

➡ CompositionCountそれ自体に眼球運動データを説明できる効果がある

ALL:

```
RT ~ [baselineの回帰式] + CCright + CCleft + CCreveal
```

結果・考察

2. 右枝分かれ vs. 左枝分かれ

	χ^2	df	p
Baseline < Right	91.438	1	< **0.0001**
Baseline < Left	126.96	1	< **0.0001**
Left < RightLeft	0.5483	1	0.459
Right < RightLeft	36.068	1	< **0.0001**

➡ 眼球運動データに説明力は、`CCleft > CCright`

- 日本語において逐次的に構築されていると考えられる構造として、CCGの左枝分かれ構造の方が妥当

結果・考察

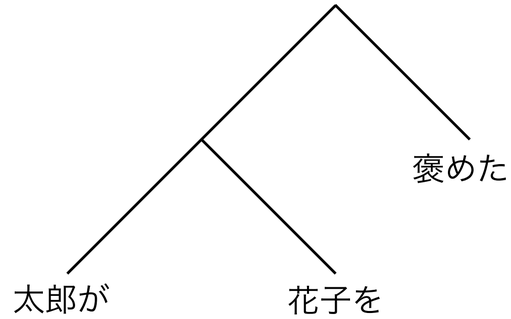
3. 左枝分かれ vs. reveal操作

	χ^2	df	p
Baseline < Left	126.96	1	< **0.0001**
Baseline < Reveal	125.72	1	< **0.0001**
Reveal < LeftReveal	1.2392	1	0.2656
Left < LeftReveal	0.0004	1	0.9837

➡ 眼球運動データに説明力は、$\texttt{CCleft} \not< \texttt{CCreveal}$

日本語において逐次的に構築されていると考えられる構造として、reveal操作による構造の方が妥当とは言えない

まとめ



- 日本語の逐次処理で構築される構造として、
  CCGの右枝分かれ構造より左枝分かれ構造の方が妥当
  - 動詞に先んじた項関係の計算を、計算心理言語学の知見から支持

- 左枝分かれ構造よりreveal操作による構造の方が妥当とは言えない
  - reveal操作が通言語的には妥当でないことを示唆

  ➡ reveal操作を、その前提から考え直す必要性