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
{kohei-kajikawa,yoshiryo0617,oseki}@g.ecc.u-tokyo.ac.jp

概要

言語構造の違いから，ある言語で妥当だと主張されている文処理方略が，必ずしも他の言語においても妥当であるとは限らない．英語において，Combinatory Categorial Grammar の右枝分かれ構造より左枝分かれ構造の方が，さらに reveal 操作という特別な操作を導入した方が，人間の逐次的な文処理をより高い精度でモデリングできると示されているが，本研究では，同様の主張が日本語においても成立するのか検証する．結果，左枝分かれ構造は右枝分かれ構造より妥当であると示された一方で，reveal 操作が有効であるとは言えず，英語では有効だと示された reveal 操作が通言語的に有効であるとは限らないことが示唆された．

1　はじめに

CCG (Combinatory Categorial Grammar [1, 2]) は，弱文脈依存文法の一種で高い記述力を備えている上 [3, 4]，柔軟な構成素構造により左右両方の枝分かれ構造を作ることができる．特に左枝分かれ構造は逐次的な意味計算を可能にするため，人間が逐次的に構築していると考えられる構造としての妥当性が高いと主張されている [5, 6]．実際に，英語において CCG の左枝分かれ構造に基づく処理負荷の予測が，右枝分かれ構造に基づく処理負荷の予測よりも高い精度で脳活動を説明できるということが示されている [7]．さらに，reveal 操作という，CCG で逐次的に左枝分かれ構造として処理できる構造を拡張する操作 [8] を加えることで，人間の脳活動をさらに良く説明できることが示されている [7, 9]．

一方で，head-final な日本語における CCG の左枝分かれ構造は，動詞を待たずに項構造を構築することを要求する．心理言語学において，日本語では項構造が動詞に先んじて構築されるという主張があるが [10, 11, 12]，そのような主張をナチュラリスティックなコーパスを通して検証した研究はいまだなく，検証の余地がある．また，reveal 操作により逐次的な処理が可能になる構造として，英語の後置修飾詞による修飾構造が挙げられているが [8]，日本語は原則として後置修飾がなく [13]，reveal 操作が必要な状況が少ない．さらに，関係節が埋め込まれている文は，日本語においても reveal 操作を仮定することで逐次的な処理が可能となるが，そのような文は構造の再解釈を要する典型的なガーデンパス文であるとされている [14]．こうしたことから，英語では有効だと示された reveal 操作が通言語的に有効であるとは限らないと考えられる．

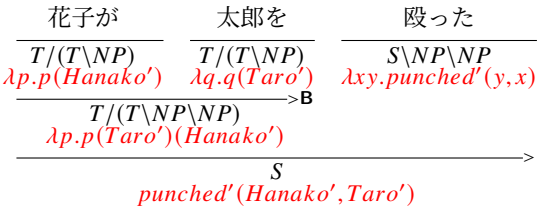


図 1　CCG の左枝分かれ構造の導出例．統語カテゴリは戸次 [2] に従い，T は変数を表す．赤文字では各構成素の意味を表している．

そこで本研究では，日本語においても，逐次的に構築していると考えられる構造として，CCG の左枝分かれ構造は右枝分かれ構造よりも妥当であるか，さらに，reveal 操作を加えることでより妥当性が向上するのかをそれぞれ眼球運動データ [15] を通して比較・検証する．

2　CCG による逐次的な文処理

2.1　左枝分かれ構造の構築

CCG は，その柔軟な構成素構造から，通常の句構造文法では右枝分かれ構造が想定される文を左枝分かれ構造で表現することができる．特に，日本語のような動詞の項が動詞に先行する言語においても，項同士の関係を先に計算することができる (図 1)．

人間の文処理は，統語処理においても意味処理においても逐次的であるが (e.g., [10, 16])，CCG での構

成素構造が実際の文処理においても構築されている場合 [1, 17, 18]，CCG の左枝分かれ構造を構築するモデルは，逐次的な統語と意味の計算が実現でき，人間の逐次的な文処理を説明できると考えられている [5, 6]．また心理言語学において，動詞の項が動詞に先行する日本語では，動詞を待たずに項構造が計算されていることが示唆されているが [10, 11, 12]，それらは特定の構造においてのみ示されており，構造に依存しないものであるかは示されていない．

2.2 Reveal 操作

一般に，必須項ではない後置修飾詞の存在は，逐次的な文処理の実現における障害となる [19]．例えば (1) のような文を想定したとき，後置修飾詞である *quickly* は，主語より低い位置で動詞句に付加するためその存在をあらかじめ予測しておく必要があるが，必須項でないため確実に予測しておくことはできない．

(1) Mary reads papers quickly.

CCG において，逐次的な文処理を実現するためには左枝分かれ構造が必要である一方，後置修飾詞は右枝分かれ構造を要求する．両者を満たすため，reveal 操作 [8] は，左枝分かれ構造を構築しながら，可能ならば左枝分かれ構造を右枝分かれ構造に変換し (木の回転)，後置修飾詞は既に作った構成素に付加するという処理を行う (図 2)．木の回転は，決定的かつ意味解釈に影響を与えないため，処理負荷がかからないと仮定されており，reveal 操作は，より少ない操作での後置修飾詞の処理を可能にする．英語における reveal 操作の認知的妥当性は，脳活動データを通して示されている [7, 9]．

一方で，日本語は原則として後置修飾がないが [13]，head-final であることから項が動詞に先行するため，関係節が埋め込まれている文は CCG でも右枝分かれ構造を要求する (2)．そのため，日本語においては，関係節が埋め込まれている文に対して reveal 操作を適用することができる．[1] しかしながら，そのような文は逐次的に処理する際，節境界がどの項の間に置かれるかが一時的に曖昧になることからガーデンパス効果が生じ，特に先行詞の部分で処理に時間を要することが知られている [14]．これ

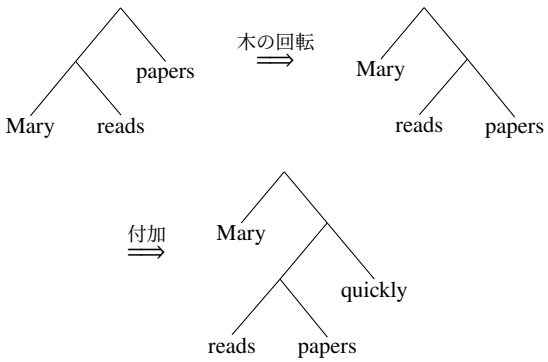


図 2 Reveal 操作による *Mary reads papers quickly* の逐次的な構造構築．右枝分かれ構造が構成できるときは，左枝分かれ構造を回転させて右枝分かれ構造を構成する．後置修飾詞はすでに作ってある右枝分かれ構造に付加する．

は，reveal 操作の予測に反しており，日本語においては reveal 操作が有効ではない可能性が指摘できる．

(2) [[花子が [[太郎を 殴った] 次郎を]] 見つけた][2]

3 実験

3.1 方法論

本研究では，以下の 2 つの仮説を検証する：日本語において，逐次的に構築していると考えられる構造として，(i) CCG の右枝分かれ構造より左枝分かれ構造が妥当である，(ii) 左枝分かれ構造より reveal 操作による構造が妥当である．

計算心理言語学において，どの文法形式，どの文処理方略が文処理理論として適切かを研究する方法として，何らかの橋渡し仮説を通して文法形式・文処理方略から得られる処理負荷を算出し，人間の行動データに対する説明力を評価するというものがある [20]．具体的には，眼球運動データや脳活動データなどの人間の行動データがアノテーションされたコーパスに対して，何らかの文法形式と文処理方略をもとに構文構造を割り当て，そこから処理コストを算出する．そしてその処理コストを含めた回帰モデルの当てはまりの良さを比較し，どの理論が最も説明力が高いかを調べるというものである [21, 22, 23, 7, 9]．そのような方法論のもと本研究では，CCG の文処理方略ごとに構築した構造で，意味合成が起こる数を橋渡し仮説として処理負荷を算出した．そして，行動データとして，文節ごとの視線

1) このほか，“太郎が学会で仙台に，旅行で札幌に行った” のような非構成素等位接続構造にも reveal 操作は適用できると考えられるが，本稿では詳述しない．

2) ブラケットで示した構成素構造は，CCG で作ることのできる最も左枝分かれな構造を示している．

走査法による読み時間がアノテーションされている BCCWJ-EyeTrack [15] を用い，読み時間に対する回帰モデルの精度比較を通して，CCG における文処理方略ごとの認知的妥当性を比較した．

3.2 木構造の獲得

CCG の右枝分かれ構造は，depccg[3] [24] により BCCWJ-EyeTrack の各文に対して割り当てた．統語カテゴリが割り当てられる単語単位が BCCWJ-EyeTrack の文節境界を跨がないように，各文節を事前に Janome[4] で分割し，その結果を depccg の入力とした．CCG の左枝分かれ構造は，depccg に解析された構造のなかで，後方関数適用 (<) により項として合成されている構成素すべてに型繰上げ規則 (>**T**) を適用したのち，可能な限り左枝分かれ構造に回転させて構成した．詳細は付録に示す．

3.3 橋渡し仮説

先行研究では，逐次的に構築されると考えられる構造と実際の行動データを繋げる橋渡し仮説として，単語ごとに構築されるノードの数を，各単語における処理負荷と仮定している [21, 22, 23, 7, 9]．特に，文法形式として CCG を用いた Stanojević ら [7, 9] は，unary なノードも含めたすべてのノードの数を処理負荷としている．しかし本研究では，CCG の構文解析器特有の変換規則や [25]，左枝分かれ構造に変換するのに過剰に適用された型繰上げ規則の影響を排して意味合成における処理負荷を明確に考慮するため，各単語ごとに新たに構築される二分木の数，すなわち，意味合成が起きる数を各単語における処理負荷と仮定する．これを *CompositionCount* と呼ぶ (図 3)．

以下，CCG の右枝分かれ構造の CompositionCount を `CCright`，左枝分かれ構造の CompositionCount を `CCleft`，reveal 操作によって構造を構築した際の CompositionCount を `CCreveal` と略記する．

3.4 統計分析

読み時間に対する回帰モデルの精度比較のため，あらかじめ読み時間をモデリングしたベースライン回帰モデルを設定した．まず，ベースラインモデルと，ベースラインモデルに本研究で用いる 3 種類の CompositionCount すべてを説明変数として加えた回帰モデルとの尤度比検定を行い，CompositionCount 自体に眼球運動データを説明できる効果があるのか確認した．そして，ベースラインモデルに各 CompositionCount を順に説明変数として加えた回帰モデルを設定し，ネストしたモデル同士を尤度比検定で比較した．一方の説明変数が，他方の説明変数が既に説明した以上に眼球運動データを説明できるか検証することを目的としている．

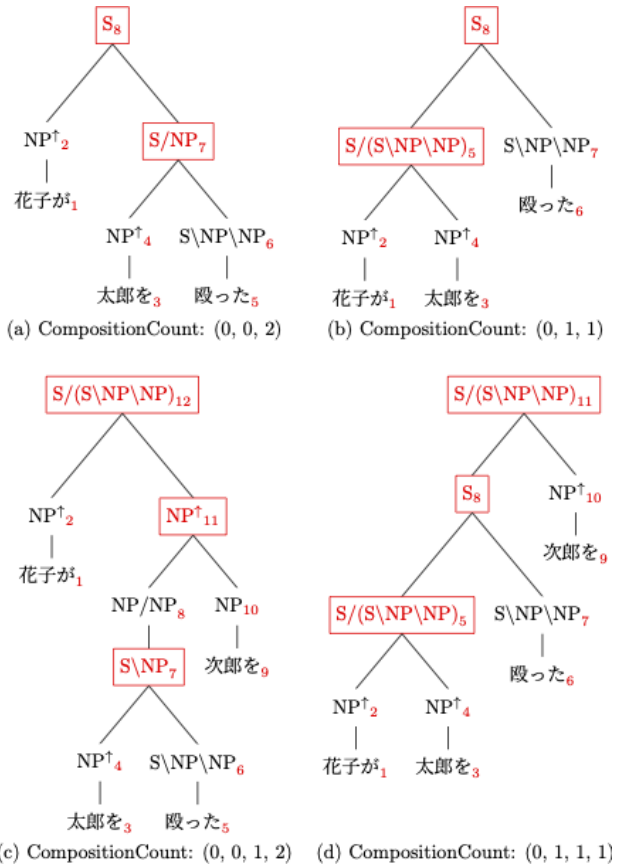


図 3 (a) 右枝分かれ構造の，(b) 左枝分かれ構造の，(c) そして両者を含む構造の CompositionCount．および，(d) reveal 操作による構築の CompositionCount．各ノードの赤い数字はそのノードが構築される順序を示す．また，赤い枠線で囲まれたノードは意味合成が起こるノードを表す．

ベースライン回帰モデルは，以下の式である：

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

各説明変数の詳細は付録に示す．読み時間として，視線走査法により計測された総注視時間 (total time) を用いた．浅原ら [15] に従い，総注視時間がゼロミリ秒である文節および本文でない文節は除外した．結果として 19,176 文節中 13,232 の文節を扱った．

3) https://github.com/masashi-y/depccg
4) https://github.com/mocobeta/janome

すべてのモデルは，article と subject のランダム切片を含んでいる．尤度比検定の結果は，$\alpha = 0.05/t$ の閾値で統計的に有意とみなした．t はボンフェローニ補正に従い，検定数である．

4 結果

以下，Baseline は 3.4 項で導入したベースライン回帰モデルを，Right, Left, Reveal, RightLeft, LeftReveal, All はそれぞれ，ベースライン回帰モデルに CCright, CCleft, CCreveal, CCright と CCleft の両方，CCleft と CCreveal の両方，CCright, CCleft, CCreveal の 3 種類全て，を加えた回帰モデルを指す．

4.1 CompositionCount の有効性

まず，Baseline と All の尤度比検定 ($\alpha = 0.05$) を行い，CompositionCount の眼球運動データに対する有効性を検証した．結果は以下，表 1 に示す：

表 1 Baseline と All の間の尤度比検定の結果．

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

検定の結果，Baseline と All には有意に差があることから，CompositionCount それ自体に眼球運動データを説明できる効果があることが確認された．

4.2 右枝分かれ構造と左枝分かれ構造

眼球運動データに対する，CCright と CCleft の説明力を nested model comparison ($\alpha = 0.125 = 0.05/4$) で比較した．結果は以下，表 2 のとおりである：

表 2 CCright, CCleft 間の nested model comparison の結果．

	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** ***

CCleft は，Baseline に加えられたときも，Right に加えられたときも，ともに回帰モデルのあてはまりの向上に有意に寄与した一方，CCright は Baseline に加えられたときのみ有意に回帰モデルのあてはまりを向上させた．これらのことから，CCleft は，視線計測データについて CCright が説明できる全てのことが説明できていると言える．つまり，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 の右枝分かれ構造より左枝分かれ構造が妥当であることを示唆している．

4.3 左枝分かれ構造と reveal 操作

眼球運動データに対する，CCleft と CCreveal の説明力を nested model comparison ($\alpha = 0.125 = 0.05/4$) で比較した．結果は以下，表 3 のとおりである：

表 3 CCleft, CCreveal 間の nested model comparison の結果．

	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

CCleft と CCreveal はともに，Baseline に加えられると有意に回帰モデルのあてはまりを向上させた一方，それぞれ Reveal，Left に加えられても，ともに有意にははたらかなかった．つまり，日本語においては，reveal 操作による構造構築が左枝分かれ構造を構築するより妥当であるとは言えない．この結果は，reveal 操作が通言語的に有効であるとは限らないことを示唆する．

5 おわりに

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も左枝分かれ構造の方が妥当であることが示された．日本語における左枝分かれ構造は，動詞に先んじた項構造の計算を要求するので，この結果は，日本語の項構造が動詞の前で構築されるという主張 [10, 11, 12] が，計算心理言語学的な知見からも支持されることを示している．また，英語では reveal 操作が有効であるという結論が導かれた Stanojević ら [7, 9] の結果と異なり，日本語においては，その reveal 操作がとりわけ妥当であるとは言えなかった．これは，言語の構造に起因していると考えられ，reveal 操作が通言語的に有効であるとは限らないことを示唆している．

一方で，reveal 操作は，その前提を変えることで，日本語における再解釈の過程を段階的に説明しうる非常に魅力的な操作である．今後は，本研究の結果を踏まえ，Stanojević らによる reveal 操作をその前提から見直し，英語と日本語双方に矛盾のない操作に改良し得るか検討した上で，逐次的な文処理方略の解明を目指していきたい．

謝辞

本研究は，JST さきがけ JPMJPR21C2 の支援を受けたものです．また本研究にあたり，多くの助言をくださった磯野真之介氏に深く感謝します．

参考文献

[1] Mark Steedman. **The syntactic process**. MIT press, 2000.

[2] 戸次大介. 日本語文法の形式理論. くろしお出版, 2010.

[3] Aravind K. Joshi. Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? In David R. Dowty, Lauri Karttunen, and Arnold M. Zwicky, editors, **Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives**, Studies in Natural Language Processing, pp. 206–250. Cambridge University Press, 1985.

[4] Edward P. Stabler. The epicenter of linguistic behavior. **Language down the garden path: The cognitive and biological basis of linguistic structures**, pp. 316–323, 2013.

[5] Mark Steedman and Jason Baldridge. Combinatory Categorial Grammar. In Robert D. Borsley and Kersti Börjars, editors, **Non-Transformational Syntax: Formal and Explicit Models of Grammar**, pp. 181–224. Wiley-Blackwell, 2011.

[6] Mark Steedman. **Taking scope: The natural semantics of quantifiers**. MIT Press, Cambridge, MA, 2012.

[7] Miloš Stanojević, Shohini Bhattasali, Donald Dunagan, Campanelli Luca, Mark Steedman, Jonathan Brennan, and John Hale. Modeling incremental language comprehension in the brain with Combinatory Categorial Grammar. **Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics**, pp. 23–28, 2021.

[8] Miloš Stanojević and Mark Steedman. CCG parsing algorithm with incremental tree rotation. **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 228–239, 2019.

[9] Miloš Stanojević, Jonathan R. Brennan, Donald Dunagan, Mark Steedman, and John T. Hale. Modeling structure-building in the brain with CCG parsing and large language models. **arXiv preprint arXiv:2210.16147**, 2022.

[10] Yuki Kamide and Don C. Mitchell. Incremental pre-head attachment in Japanese parsing. **Language and Cognitive Processes**, Vol. 14, No. 5-6, pp. 631–662, 1999.

[11] Edson T. Miyamoto. Case markers as clause boundary inducers in Japanese. **Journal of Psycholinguistic Research**, Vol. 31, pp. 307–347, 2002.

[12] Shinnosuke Isono and Yuki Hirose. Locality effect before the verb as evidence of pre-verb reactivation. **The Japanese Society for Language Sciences 23rd Annual International Conference**, 2022.

[13] Joseph H. Greenberg. **Universals of language**. MIT press, 1963.

[14] 井上雅勝. 構造的曖昧文の理解におけるガーデンパス化：眼球運動データを指標として. 日本教育心理学会総会発表論文集, Vol. 32, p. 378, 1990.

[15] 浅原正幸, 小野創, 宮本エジソン正. BCCWJ-EyeTrack. 言語研究, Vol. 156, pp. 67–96, 2019.

[16] Yuki Kamide, Christoph Scheepers, and Gerry T. M. Altmann. Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. Vol. 32, pp. 37–55. Springer, 2003.

[17] Ivan A. Sag and Thomas Wasow. Performance-compatible competence grammar. In Robert D. Borsley and Kersti Börjars, editors, **Non-transformational syntax: Formal and explicit models of grammar**, pp. 359–377. Wiley-Blackwell, 2011.

[18] Shevaun Lewis and Colin Phillips. Aligning grammatical theories and language processing models. **Journal of Psycholinguistic Research**, Vol. 44, No. 1, pp. 27–46, 2015.

[19] John T Hale. **Automaton theories of human sentence comprehension**. Center for the Study of Language and Information, 2014.

[20] Jonathan Brennan. Naturalistic sentence comprehension in the brain. **Language and Linguistics Compass**, Vol. 10, No. 7, pp. 299–313, 2016.

[21] Jonathan Brennan, Yuval Nir, Uri Hasson, Rafael Malach, David J Heeger, and Liina Pylkkänen. Syntactic structure building in the anterior temporal lobe during natural story listening. **Brain and language**, Vol. 120, No. 2, pp. 163–173, 2012.

[22] Jonathan R. Brennan, Edward P. Stabler, Sarah E. Van Wagenen, Wen-Ming Luh, and John T. Hale. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. **Brain and Language**, Vol. 157-158, pp. 81–94, 2016.

[23] Jixing Li and John Hale. **Grammatical predictors for fMRI time-courses**. Oxford University Press, 2019.

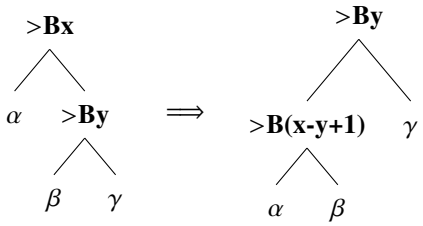
[24] Masashi Yoshikawa, Hiroshi Noji, and Yuji Matsumoto. A* CCG parsing with a supertag and dependency factored model. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 277–287. Association for Computational Linguistics, 2017.

[25] 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 Linguistics (Volume 1: Long Papers)**, pp. 1042–1051. Association for Computational Linguistics, 2013.

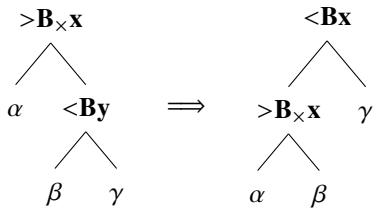
A　木の回転

本研究で，左枝分かれ構造を構成するのに用いた木の回転操作の条件を示す．以下 [8] に従い，組合せ規則を一般化して，その規則の階数 (項の数) を付して表記する．例えば，前方関数適用 (>) は **>B0**，前方関数合成 (**>B**) は **>B1** と表記する．また，X/X または $X\backslash X$ の形をした統語カテゴリを修飾語と呼ぶ．α, β, γ は構成素の統語カテゴリを表し，**x**, **y** は変数 (整数) を表す．

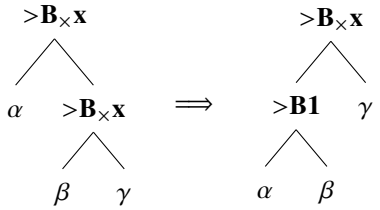
1. **x** ≥ **y** のとき[5)]



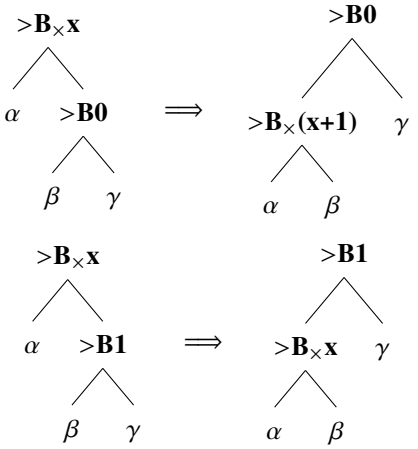
2. 1 を満たさず，かつ修飾語を含んでいるとき
   (a) α, γ がともに修飾語であるとき



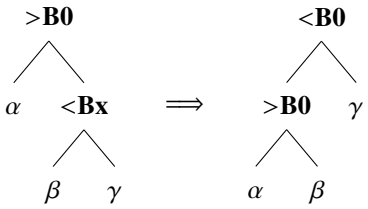
   (b) α, β がともに修飾語であるとき



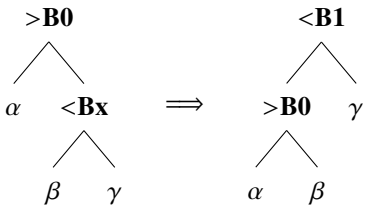
   (c) α が修飾語であるとき



   (d) γ が修飾語であるとき
      i. α が，$S/(S\backslash NP)$ のような，複合カテゴリでありかつその左側カテゴリが複合カテゴリでないとき



      ii. α が，$(S\backslash NP)/(S\backslash NP\backslash NP)$ のような，複合カテゴリでありかつその左側カテゴリも複合カテゴリであるとき



これらの回転操作は，BCCWJ-EyeTrack に対する depccg の解析結果が最も左枝分かれになるように構成したものである．特に 2 以降の条件については，本研究で用いられた統語カテゴリ以外の統語カテゴリが含まれている構造に対して，必ずしも有効ではない．

B　説明変数

本研究において，ベースライン回帰モデルで用いた説明変数を表 4 に示す．

表 4　本研究で用いた説明変数

変数名	型	記述
RT	int	読み時間
dependent	int	係り受け関係
length	int	文字数
frequency	num	文節の単語頻度の幾何平均
is_first	factor	行内最左要素
is_last	factor	行内最右要素
is_second_last	factor	行内右から 2 番目の要素
screenN	int	画面提示順
lineN	int	行提示順
segmentN	int	文節提示順
article	factor	記事番号
subject	factor	被験者番号

5)　[8] では **x** > **y** と表記されているが，**x** ≥ **y** の誤りと思われる．反例は，$\alpha = S/(S\backslash NP), \beta = S\backslash NP/NP, \gamma = NP$, **x** = **y** = 0 のときである．