タグおよびアノテーションの概要

統語・意味解析コーパス (NPCMJ) チュートリアル（2）

吉本啓

2021.3.13

はじめに

- NPCMJ のアノテーション（タグおよび構文分析）の基本
  木構造のパターンを使った例文の検索のために必須

NPCMJ の出典（2021 年 3 月現在）

Source	ツリー数	語数	
aozora	12,809	246,708	青空文庫
bible	1,664	26,119	聖書
blog	219	3,217	ブログ
book	553	10,992	書籍
dict	26,279	141,201	辞書
diet	1,698	32,446	国会会議録
essay	541	11,502	エッセイ
fiction	958	10,445	フィクション
law	337	6,954	法律文
misc	2,211	22,745	その他
news	5,981	90,137	ニュース
nonfiction	234	4,124	ノンフィクション
spoken	2,382	12,578	会話
ted	1,453	21,366	TED トーク
textbook	6,953	63,974	教科書
wikipedia	2,746	59,758	ウィキペディア
Total	67,018	764,266	

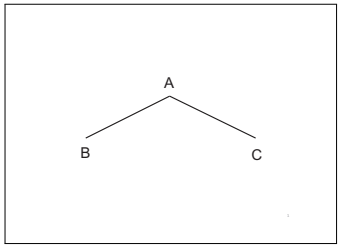
タグ（ラベル）付けの一般規則

カッコ表示と木表示

句構造規則の適用の結果生じる派生の過程（「解析結果」とも呼ぶ）を図示したもの

A ⇒ B C

(A
B
C)

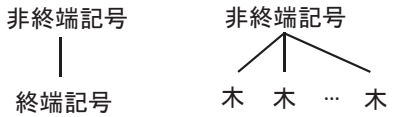


どちらでも情報は全く同じ

カッコ表示の行換えは無意味

表示の規則

- 木 → (非終端記号　　終端記号)
- 木 → (非終端記号　木
  木
  …
  木)



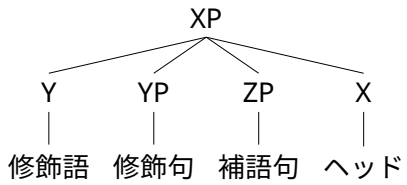
X バー理論 (1)

句や節の種類 (文，動詞句，名詞句，形容詞句等) の区別に関わらず，主部，補部，修飾部等，内部構成素の機能や配列は並列的

- The enemy destroyed the city
- The enemy's destruction of the city

X バー理論 (2)

- 抽象的な文法スキーマ：
  - 文法規則を作る規則
  - X は XP を投射（project）する



X バー理論 (3)

日本語の場合，語順も共通

[[汽車で] 修飾語 [記者が] 補語句 [帰社する] ヘッド] S
[[とても] 修飾語 [鯛焼きが] 補語句 [好きだ] ヘッド] AdjP
[[ハンサムな] 修飾語 [山田君の] 補語句 [お父さん] ヘッド] NP

課題 1

以下の統辞構造を木で示し，それぞれどの部分が修飾語・句，補語句およびヘッドに相当するか説明しなさい。

- 幼少期の一休の名前
- 一番頭がいい
- よく映画を見る

動詞やイ形容詞の活用形の扱い

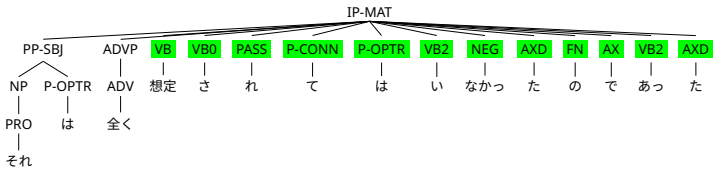
動詞，イ形容詞，助動詞などの活用語に関しては，概ね学校文法的な扱いを採用している

- 動詞（VB）
  - 下一段活用：「食べる」→ 食べる｜食べろ｜食べよ｜食べ｜食べれ
  - 上一段活用：「起きる」→ 起きる｜起きろ｜起きよ｜起き｜起きれ
  - 五段活用：「走る」→ 走る｜走れ｜走ら｜走り｜走ろ
  - カ変活用：「来る」→ 来る｜来い｜来れ｜来
  - サ変活用：「する」→ する｜しろ｜せよ｜すれ｜し

述語の拡張

- IP (節) を投射するのは述語
- 拡張された述語の個々の要素は，同一の節（IP）の元にフラットに並べられる

「それは全く [想定されてはいなかったのであった]」



様々なタグ（1）

- タグのタイプには次のようなものがある
  - 「語レベルのタグ」と「句レベルのタグ」
  - 「基盤となるタグ」と「拡張タグ」

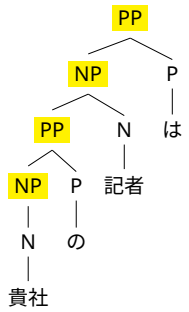
- 品詞タグ：語レベルのタグ
  - 語に与えられる

    N（Noun，名詞），P（Particle，助詞），VB（Verb，動詞），ADJI（I-adjective，イ形容詞），ADV（Adverb，副詞），...

様々なタグ（2）

- 統語タグ：句レベルのタグ
  - （多くの場合）語レベルのカテゴリーが投射するカテゴリーに対して与えられる

    NP（Noun Phrase，名詞句），PP（Particle Phrase，助詞句），ADVP（Adverbial Phrase，副詞句），IP（Inflectional Phrase，節）...



- NP → N
  (NP (N 貴社))
- PP → NP P
  (PP (NP (N 貴社))
  (P の))
- NP → PP N
  (NP (PP (NP (N 貴社))
  (P の))
  (N 記者))

様々なタグ (3)

- 基盤タグ：
  - 語やそれの投射するカテゴリーに対して与えられるタグ
- 拡張タグ：
  - 語または句の下位類や統語的な機能を示すタグ
  - 基盤タグの後にハイフンを付けて付加される

    IP-MAT（Matrix，主節），PP-SBJ（Subject，主語である助詞句），P-ROLE（格助詞），P-OPTR（とりたて助詞），NP-PRD（Predicate，名詞述語）

品詞タグ (1)

名詞類

名詞	N	海，平和，木
固有名詞	NPR	大阪，山手線，鈴木さん，東北大学
形式名詞	FN	はず，よう，もの，の
代名詞	PRO	私，彼女，これ，そちら
疑問代名詞	WPRO	何，だれ，どこ，どちら
数詞	NUM	一，150，約一億二千万

品詞タグ (2)

動詞類

動詞（語幹）	VB	歩く，食べる，急ぐ，なさる
軽動詞	VB0	する，できる，いたす，願う
補助動詞	VB2	いる，おく，くれる，なさい
助動詞	AX	だ，たい，らしい，そう（だ），ます，（よ）う
テンス指標	AXD	た
否定辞	NEG	ない，ん，ず，まい，な
受動助動詞	PASS	（ら）れる
間接受動助動詞	PASS2	（ら）れる

品詞タグ (3)

モーダル要素

MD	かもしれない，そう，だろう，ちがいない，であろう，でしょう，なければならない，べき，みたい，よう，らしい，わけ，相違ない
ADJI-MD	いい，ない，なかろう，よい，よろしい，難く
ADJN-MD	だめ，結構

品詞タグ (4)

形容詞類

イ形容詞	ADJI	美しい，長い
ナ形容詞	ADJN	親切だ，簡単だ，決然たる，呆然と

（タル・ト型を含む）

疑問詞類

疑問副詞	WADV	どう，どうして，いかに，なぜ
疑問限定詞	WD	どの，どんな，どういう，いかなる
疑問数詞	WNUM	何，何百，いくら
疑問代名詞	WPRO	何，だれ，どれ，どちら，いつ，どこ

品詞タグ (5)

助詞

補文助詞	P-COMP	という，との，なんて
接続助詞	P-CONN	と，か，や，も，あるいは，かつ
終助詞	P-FINAL	か，ね，よ，さ
間投助詞	P-INTJ	さ，な，ね，よ
とりたて助詞	P-OPTR	か，くらい，は，も，しか，さえ
格助詞	P-ROLE	が，に，を，で，から

品詞タグ (6)

その他

副詞	ADV	あえて，ちょっと，ほとんど，もう
助数詞	CL	つ，人，メートル，年生
等位接続詞	CONJ	しかし，けれども，さて，ちなみに
限定詞	D	この，あんな，そういう
間投詞	INTJ	はい，ええ，ああ，うわあ
連体詞	PNL	大きな，ちょっとした，たいした，いろんな
量化詞	Q	みんな，たくさん，少し，全員

統語タグ (1)

名詞句 NP : 多くの場合，必須文法役割等の機能を表す拡張タグを伴う

主語名詞句	NP-SBJ
第一目的語名詞句	NP-OB1
主題名詞句	NP-TPC
時間名詞句	NP-TMP
述語名詞句	NP-PRD

統語タグ (2)

助詞句 PP : 機能を表す拡張タグを伴うことが多い

主語助詞句	PP-SBJ
副詞的助詞句	PP-ADV
場所助詞句	PP-LOC
数量助詞句	PP-MSR

統語タグ (3)

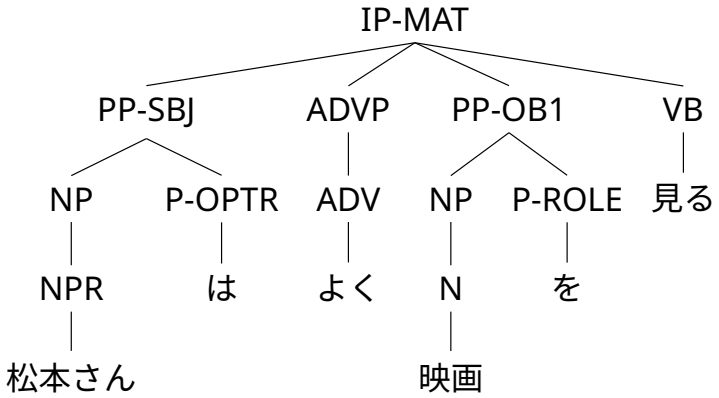
節 CP, IP : 機能を表す拡張タグを伴う

命令節	CP-IMP
疑問節	CP-QUE
補部節	CP-THT
主節	IP-MAT
副詞節	IP-ADV
関係節	IP-REL
空所なし名詞修飾節	IP-EMB

基本的な文の構成 (1)

主要文法役割：	PP または NP の拡張タグとして，SBJ（主語）， OB1 (第一目的語)，OB2（第二目的語）を表示
第一目的語：	2 項述語の，主語でない方の項 大部分が「直接目的語」と一致するが，同一ではない。
3 項述語の必須文法役割：	「を」により表示された項が OB1。 主語以外の残る項が OB2。

基本的な文の構成 (2)

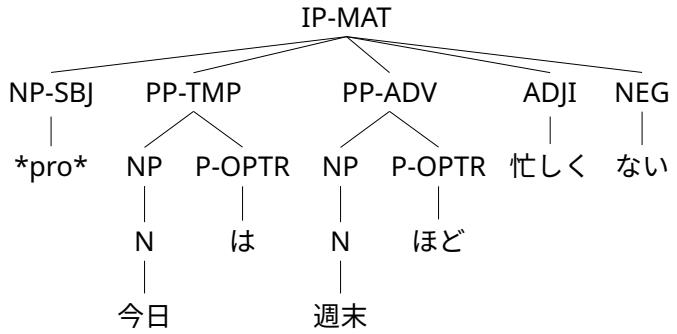


基本的な文の構成 (3)

任意文法役割： 格助詞「に」「へ」「で」「から」「まで」「と」等により作られた PP が付加詞として用いられることにより表示

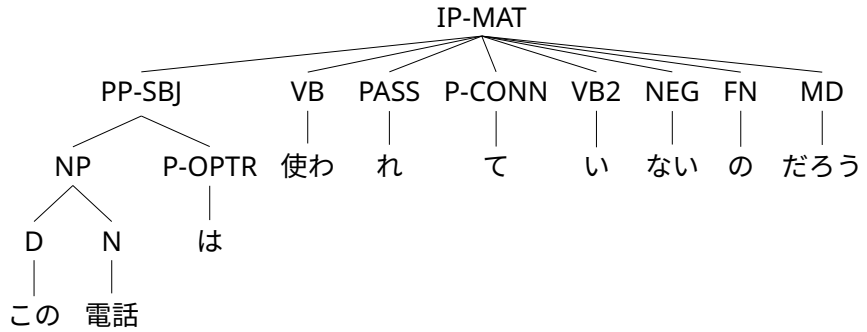
PP に LOC（場所），TMP（時間），MSR（時間軸上の範囲または頻度），ADV（その他の副詞的意味）のような拡張タグを加える（作業中）。

基本的な文の構成 (4)



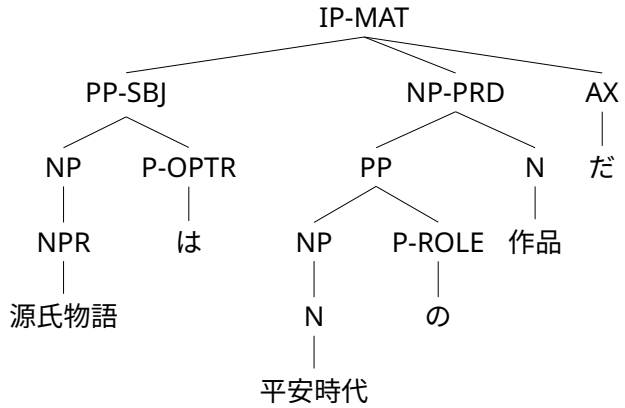
基本的な文の構成 (5)

述語拡張形を作る，接続助詞 P-CONN，とりたて助詞 P-OPTR，軽動詞 VB0，補助動詞 VB2，助動詞 AX，モーダル要素 MD，形式名詞 FN 等はすべて IP に直接支配され，フラットな構造を作る。



基本的な文の構成 (6)

コピュラをともなって述語となる名詞句は NP-PRD とする。



空要素 (1)

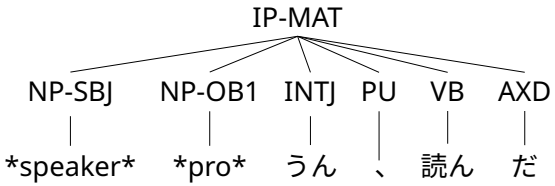
インデクスを使用しない空要素

- ゼロ代名詞
- 虚辞（主語を持たない文）
- 関係節のトレース

空要素 (2)

ゼロ代名詞

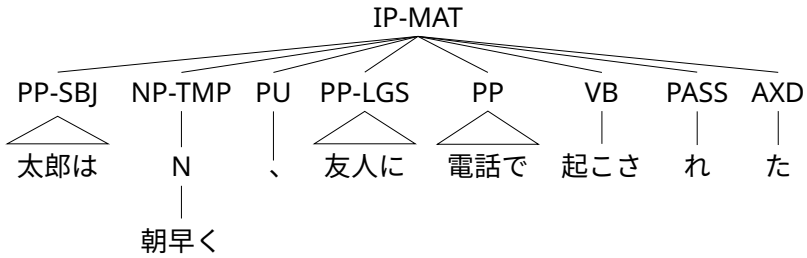
- 必須文法役割を担う省略された NP で，文脈中や文中の先行詞と同一指示関係にあるもの
- 一般には *pro* で表す。話し手や聞き手を指示する場合は，*speaker* や *hearer* とする



複雑な述語 (1)

直接受動文

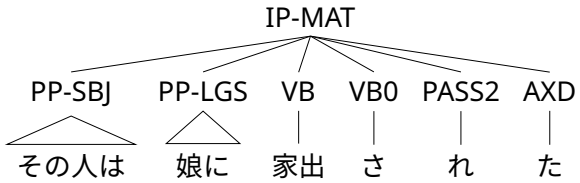
主体の意味役割を持つ助詞句：PP-LGS
受動助動詞：PASS



複雑な述語 (2)

間接受動文

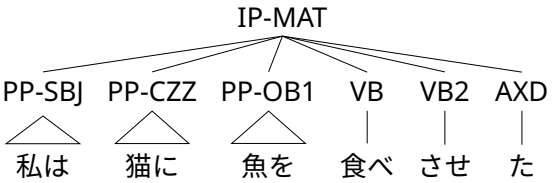
主体の意味役割を持つ助詞句：PP-LGS
受動助動詞：PASS2



複雑な述語 (3)

使役文

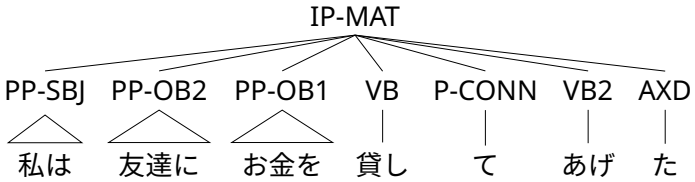
被使役者助詞句：PP-CZZ
使役助動詞：VB2



複雑な述語 (4)

てやる文
てくれる文

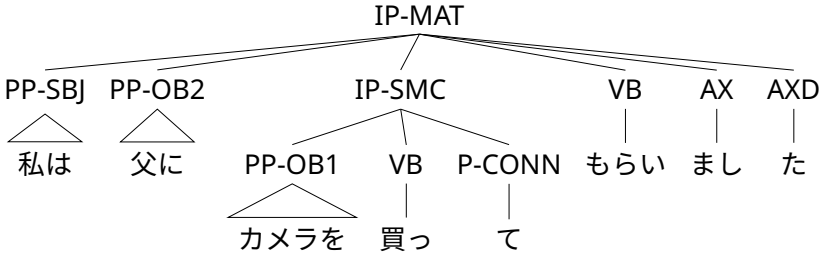
やる（あげる），くれる：VB2



複雑な述語 (5)

てもらう文

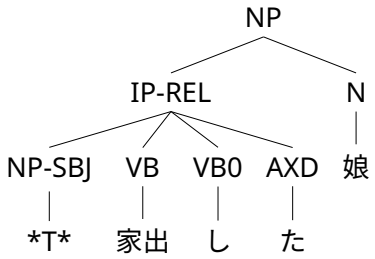
もらう：VB として扱う。
動詞句に相当する IP-SMC を埋め込みとして持つ。



関係節 - 内の関係

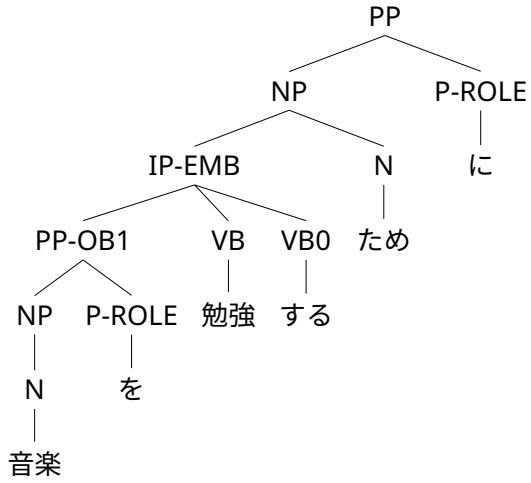
関係節：IP-REL で表す

先頭に格役割の情報を伴うトレースを置く
名詞を修飾する形容詞も関係節として扱う



関係節 - 外の関係

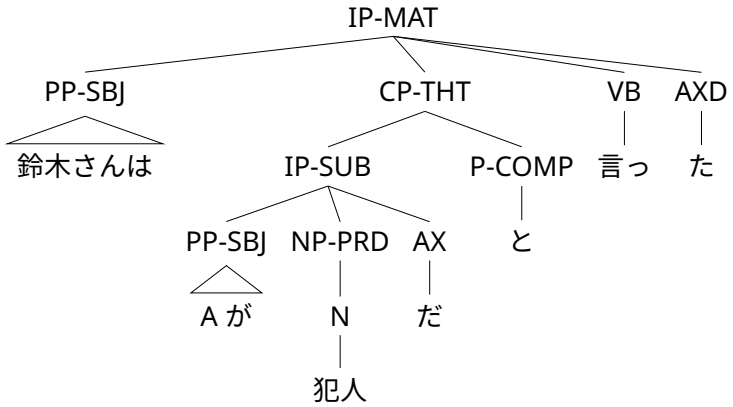
空所なし名詞修飾節：IP-EMB で表す



補部節

補部節 CP-THT:

「と」等の補文助詞 P-COMP を伴い，伝達動詞や認識動詞の補部となる。



複文 (1)

副詞節

従属節

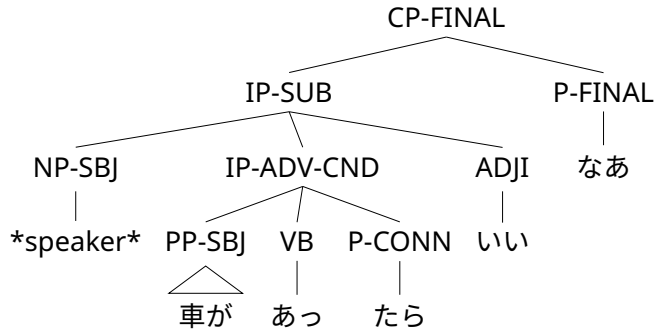
- 条件節：IP-ADV-CND，PP-CND
- それ以外の従属節：IP-ADV-SCON，PP-SCON

等位節（並列，対比，または継起）

IP-ADV-CONJ, PP-CONJ

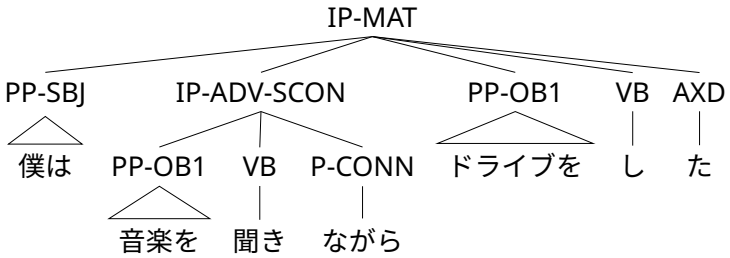
複文 (2)

条件節 IP-ADV-CND



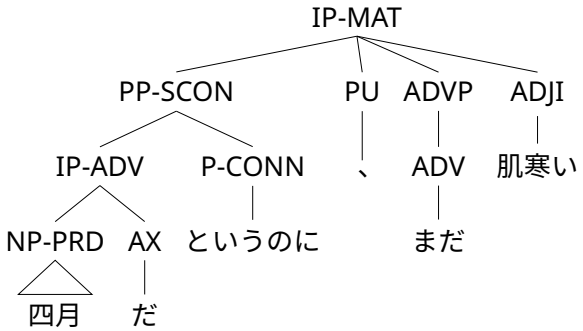
複文 (3)

条件節以外の従属節 IP-ADV-SCON



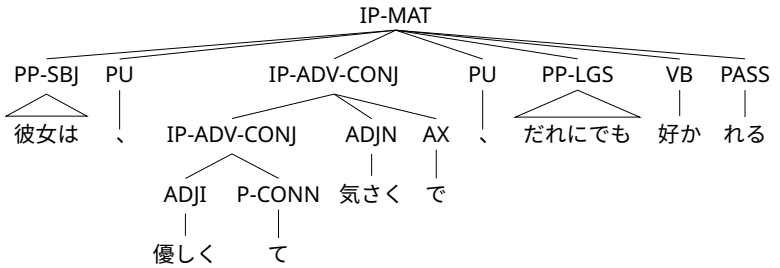
複文 (4)

条件節以外の従属節 PP-SCON



複文 (5)

等位節



コントロール (1)

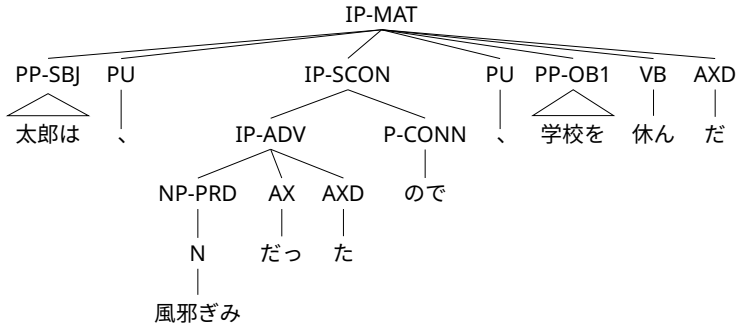
ある種の従属節（IP-ADV，IP-SUB 等）や空所なし名詞修飾節（IP-EMB）で，明示されていない主語はディフォールトとしてそれよりもすぐ上位の節の主語や第一・第二目的語と一致するものとする

- ゼロ代名詞としてのアノテーションを行わない

  意味解析により照応関係を同定
- 従属節・名詞修飾節の種類による違い
  - 主語・目的語のどれが先行詞となりうるか
  - 先行詞と従属節との前後関係
- 先行詞としてのアクセス可能性に順位

  OB2 ⟨ OB1 ⟨ SBJ2 ⟨ SBJ

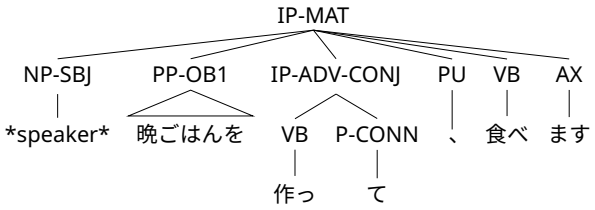
コントロール (2)



ATB 抽出

Across the Board (ATB) 抽出：等位節において行う

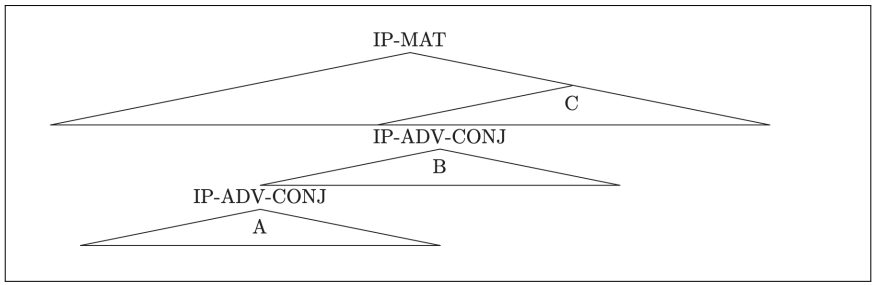
- ゼロ代名詞のアノテーションを行わない
- 埋め込まれる等位節の主語以外の項も継承される
- 上位節における項の文法役割がそのまま継承される



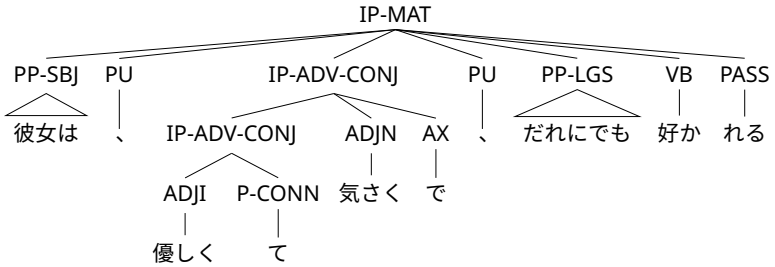
並列 (1)

並列句

- 副詞節 (IP-ADV)：副詞節がそのすぐ上位の節に再帰的に埋め込まれる



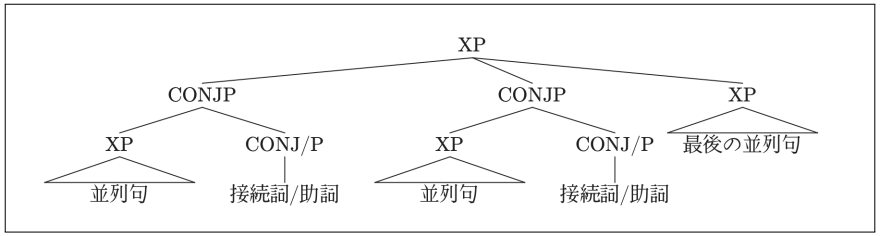
並列 (2)



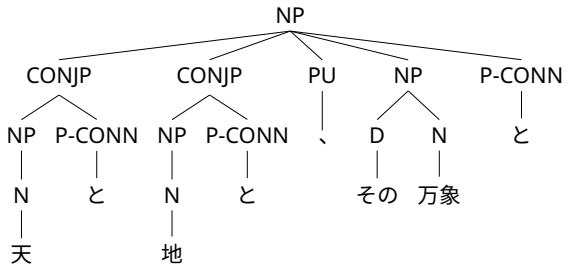
並列 (3)

並列句

- 副詞句 (ADVP)，助詞句（PP)，名詞句（NP)，疑問節（CP-QUE）等：フラットな構造を取る



並列 (4)



おわりに

- 検索したい表現の構文パターンを知る必要
  - アノテーションの概略の理解
- ヒント
  - マニュアル（HP 上にあり）に目を通す
  - 益岡・田窪『基礎日本語文法』例文のアノテーションに目を通す