

2. データの概要とタグの検索（補足資料）

1 はじめに

- NPCMJ のデータはアノテーションについてある程度理解しなければ、有効に利用できないという側面をもつ。
- 以下、いくつかの重要と考えられるアノテーションの特徴を挙げる。

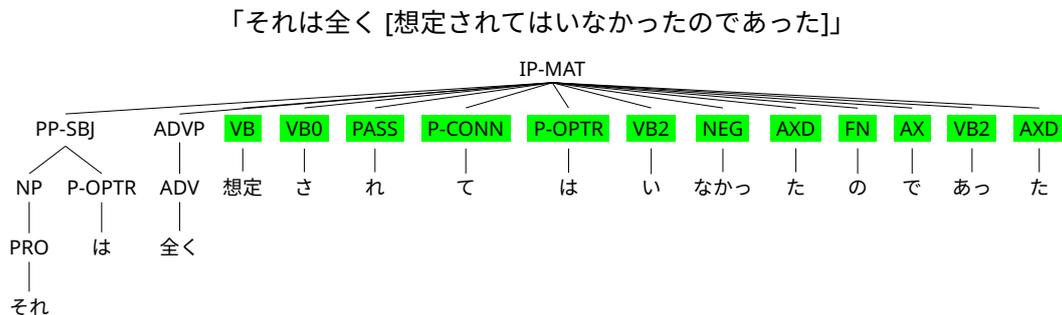
2 動詞やイ形容詞の活用形の扱い

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|食べる|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
 - 力変活用：「来る」→ 来る|来い|来れ|来
 - サ変活用：「する」→ する|しろ|せよ|すれ|し
 - 使役：「せる・させる」→ せる|せ|せれ|せよ|せろ|させる|させ|させれ|させよ|させろ
- イ形容詞（ADJI）
 - 「暑い」→ 暑い|暑かる|暑かつ|暑く|暑う|暑けれ|暑かれ
- 補助動詞（VB2・PASS）
 - 受け身（PASS），自発・可能・尊敬（VB2）→ られる|られ|れる|れ
- 助動詞（AX）
 - 願望：「たい」→ たい|たかる|たかつ|たく|たけれ
 - 丁寧：「ます」→ ます|ませ|ましよ|まし
 - コピュラ：「だ」→ だ|で|の|な|に|と
- 否定の助動詞（NEG）
 - 否定：「ない」→ ない|なかる|なかつ|なく|なけれ
(イ形容詞（ADJI）の「ない」も同様)
- テンス標識の助動詞（AXD）
 - 過去：「た・だ」→ た|たろ|だ|だろ
(ただし、「たら・だら」は P-CONN（接続助詞）として扱っている)

3 述語の拡張

述語は、動詞 (VB)、イ形容詞 (ADJI)、ナ形容詞 (ADJN)、名詞句述語 (NP-PRD)、副詞句述語 (ADVP-PRD)、助詞句述語 (PP-PRD)、連体詞 (PNL) などの核となる要素に他の要素が後続し拡張される。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。



述語を拡張する要素として、軽動詞 (VB0)、補助動詞 (VB2)、助動詞 (AX)、専用のタグをもつ補助動詞・助動詞、モーダル要素 (MD)、形式名詞 (FN)、接続助詞 (P-CONN)、とりたて助詞 (P-OPTR) がある。

- 軽動詞 (VB0)
 - 「移動する」「チェックする」「ご報告する」「お見送りする」など、「動名詞+する」型の複合動詞における「する」
 - 上記「する」と同様に用いられる「いたす」「なさる」「くださる」「できる」...
- 補助動詞 (VB2)
 - 複合動詞の後項 (生産性の高いもの): 「食べ始める」「言い尽くす」など
 - 動詞テ形の後に現れ、アスペクトや受益の表現を作る動詞:
「いる」「ある」「おく」「しまう」「みる」「みせる」「いく」「くる」「いらっしゃる」「お出で」「まいる」「ご覧」「あげる」「くれる」「やる」「くださる」「頂戴」など
 - 「～である」「よくありません」の「ある」
 - 「～たりする」「～なりする」「～さえする」「～もする」などの「する」
 - 敬語形「お書きになる」、「ご覧になる」の「なる」
 - 「多すぎる」の「すぎる」
- 助動詞 (AX)
 - イ形容詞と同じ活用形をもつ派生接尾辞: 「がたい」「たい」「やすい」
 - ナ形容詞・名詞と同じ振る舞いをする派生接尾辞: 「がち」「そう」「っぱなし」
 - いわゆる「比況・例示」の助動詞: 「ごとし (ごとし・ごとき・ごとき)」
 - コピュラ: 「だ (だ・で・の・な・に・と)」「です (です・でし・でしよ)」「じゃ (じゃ・じゃあ)」「たる (たる・たら・たり・たれ)」「や (や・やん)」
- 専用のタグをもつもの補助動詞・助動詞
 - 否定 (NEG): 「食べない」「食べず」「食べません」「食べぬ」「食べまい」などの「ない」「ず」「ん」「ぬ」「まい」
 - 受け身 (PASS): 「食べられる」「飲まれる」の「られる・れる」
(直接受け身の「られる・れる」に対するタグ, 他の「られる・れる」は VB2 とタグ付け。
*この3月のアップデート以降, 間接受け身の「られる・れる」は PASS2 とタグ付けされる)
 - テンス標識 (AXD): 「食べた」「飲んだ」の「た・だ」
- モーダル要素 (MD)
 - 「かも」「かもしれない」「かもしれません」
 - 「～そうだ」の「そう」
 - 「だろう」「であろう」「でしょう」

(用言の後に現れるもの)

- 「べし (べし・べき・べく)」「らしい」「みたい」

- 「よう・う」

(「複合的な」MD に関しては、今後のアノテーションの仕方を変更する可能性が高い)

• 形式名詞 (FN)

- 「～のだ・んだ」の「の・ん」

- 「～ところだ」の「ところ」

- 「～はずだ」の「はず」

• 接続助詞 (P-CONN) (実際は動詞の屈折接尾辞であるもののみ)

- 「食べて」「飲んで」の「て・で」

- 「食べながら」の「ながら」

- 「食べつつ」の「つつ」

- 「食べれば」の「ば」

- 「食べたり」「飲んだり」の「たり・だり」

- 「食べたら」「飲んだら」の「たら・だら」

- 「食べたって」「飲んだって」の「たって・だって」

• とりたて助詞 (P-OPTR)

- 以上の要素間に現れたとりたて助詞

4 助詞関連 1

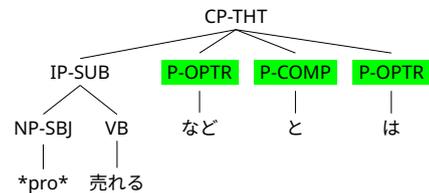
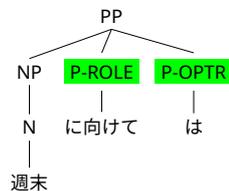
いくつかの語が緊密に連結して1つの機能語として働くものは、1つの語として扱うことを原則としている (cf. モーダル要素 (MD))。助詞 (P) にも複合的なものが数多くある。

• 複合的な助詞 (一部)

- 「をとおして」「をめぐって」「をはじめとして」「について」「にしたがって」「にわたって」「に当たって」「に向けて」「による」「として」「という」「とはいえ」「との」「だけに」「だけあって」「はもとより」...

5 助詞関連 2

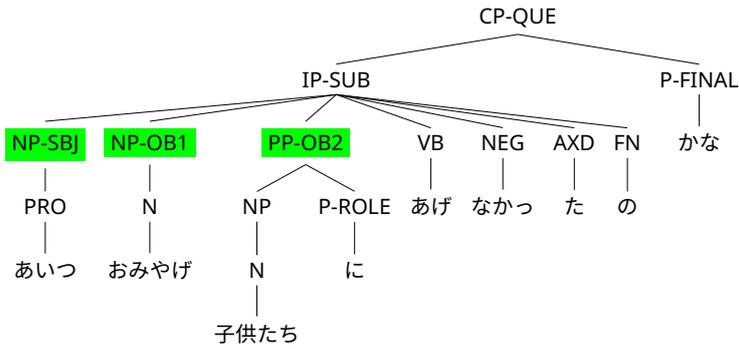
助詞 (P) が複数連続する場合は、同一の助詞句 (PP)、補部節 (CP-THT)、終助詞節 (CP-FINAL) などの下にすべてをフラットに配置することを原則とする。



6 主要文法役割

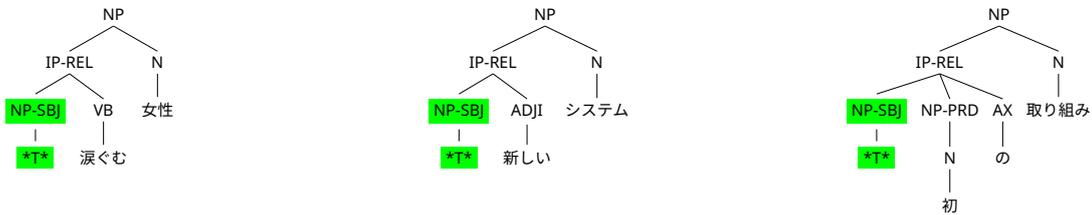
主語や目的語といった主要文法役割を果たす構成素にはそれを明示するための拡張タグを付ける：SBJ (主語)、OB1 (第一目的語)、OB2 (第二目的語)、SBJ2 (第二主語)、LGS (論理的な主語)。

これらは、構成素が助詞 (P) を伴っていればその投射する助詞句 (PP) に付加される。また、名詞句 (NP) が助詞 (P) を伴わずに主要文法役割を果たしていれば、名詞句のタグに直接付加される。

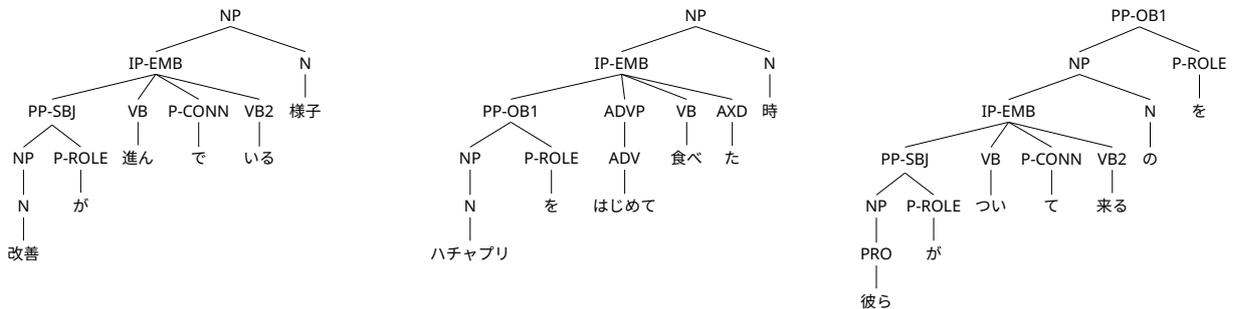


7 空要素 1

関係節 (IP-REL) 中に、主名詞 (修飾される名詞) が関係節の中で項 (主語や目的語など) や付加詞として働いていることを示すために、トレース (*T*) が置かれる。トレースが関係節の中で果たす文法役割は *T* を支配するノードに拡張タグとして示される。



なお、トレースをもたない名詞修飾節 (いわゆる「外の関係」の名詞修飾節) は、IP-REL ではなく、IP-EMB というタグが与えられる。



8 空要素 2

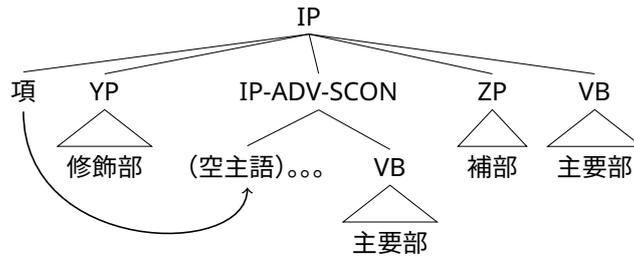
主語または目的語が省略されている場合、それらはゼロ代名詞として文中に補われる。

pro	非談話参与者
speaker	話し手
hearer	聞き手
speaker+hearer	話し手と聞き手
speaker+pro	話し手と、聞き手でない談話参与者
arb	一般的非人称指示
exp	虚辞

ただし、主語が上位の節から継承される節 (コントロール環境) および上位の節と主語や目的語などを共有する節 (ATB 環境) については、ゼロ代名詞はアノテーションはされない。

9 コントロール環境

節連続や埋め込まれた節の中の主語が、上位の節の項によって継承されるプロセスをコントロールと呼ぶ。

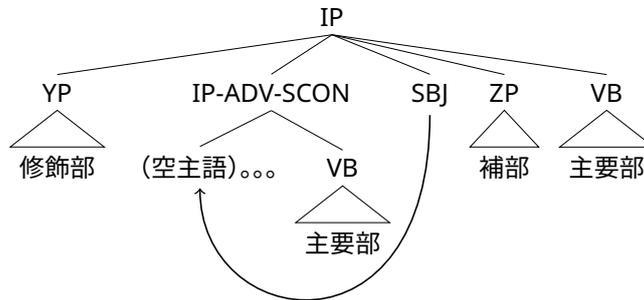


コントロール環境（コントロールが働く下位の節）として，IP-ADV-SCON（副詞節），IP-ADV-CND（条件節），IP-ADVに接続助詞が後続してできたPP-SCON，PP-CND（助詞節），IP-EMB（トレースを持たない名詞修飾節），IP-SMC（小節），IP-NML（名詞節）が想定されている。

上位の節の項は次のような階層の順に，コントロール環境中の主語となる。

OB2 > LGS > OB1 > SBJ2 > SBJ

右から左へのコントロールは，常に主語（SBJ）による。

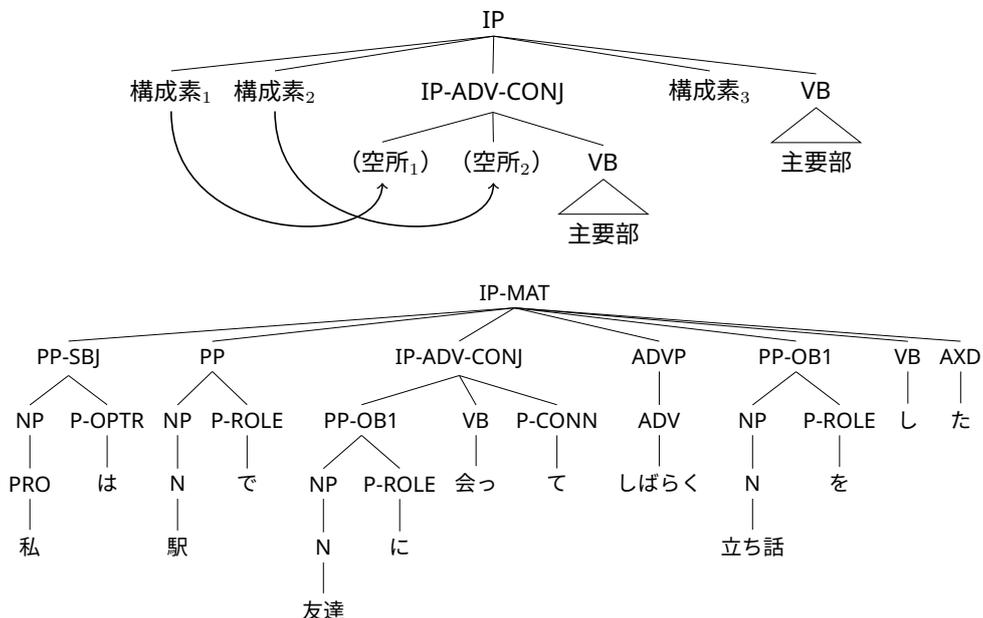


コントロール環境において継承されるゼロ代名詞主語は，アノテーションには反映されない。コントロール関係は自動的に計算される。例えば，インターフェースの default のツリーはアノテーションをそのまま木構造として示しているので，コントロール環境のゼロ代名詞主語は見えない。しかし，ツリーを indexed 表示にすると，意味計算の結果が反映され，ゼロ代名詞と照応関係が示されるようになっている。

10 ATB

等位関係にある節のうち前の節（前件）は IP-ADV-CONJ または PP-CONJ とタグづけされる。等位節における構成素の共有に関して，ATB 抽出というプロセスを想定する。

ATB 抽出：等位関係にある複数の節では，文全体の先頭，前件の外側にある構成素を，その文法役割も含めて共有する。



ATB 抽出によって共有される構成素も、ゼロ代名詞としてアノテーションされることはない。したがって、インターフェースの default のツリーではゼロ代名詞主語は見えない。ツリーを indexed 表示にすると、意味計算の結果が反映され、前件における述語の項にそれが反映される。

11 遊離した量化表現

- 遊離した量化表現を明示するために、上記のような要素を支配する NP あるいは PP のノードに、ホストと同一の文法役割の情報を ;*SBJ* や ;*OB1* のような形で付ける。

