イントロダクション

プラシャント・パルデシ
国立国語研究所
prashant@ninjal.ac.jp

1 はじめに

国立国語研究所では、第二期中期計画において『現代日本語書き言葉均衡コーパス（BCCWJ）』のデータを利用して、共起関係を主とする内容語（動詞、名詞、形容詞、副詞）の振る舞いを検索・抽出するシステム NINJAL-LWP for BCCWJ (NLB) を開発した（http: //nlb.ninjal.ac.jp/; 赤瀬川・パルデシ・今井 2016 を参照のこと）。この検索システムは、特に高度なコンピューターの知識がないユーザーでも簡単かつ瞬時に研究に必要な情報を検索・抽出・ダウンロードすることを可能にし、多数の研究者によって利用されている。この間、国内外の研究者・大学院生からコーパスに基づく機能語、句、節、複文といった様々なレベルでの構造体 (constructions) を検索・抽出できるシステムを構築してほしいとの要望を多数受けた。また、学会などでも研究者らから同様の要請を受けた。

英語に関しては、90 年代初頭から Pennsylvania（以下、Penn と略す）大学で統語解析情報タグ付きコーパス (ツリーバンク) が開発され、コーパスに基づいて機能語、句、節、複文のような様々なレベルでの構造体を大量のデータから検索・抽出して研究を行うことが可能となり、目覚ましい成果をあげている。また英語以外の世界の主要な言語について、共通の Penn 方式のフォーマットによるツリーバンクが作られ、それらの間での対照研究も活発に行われている。一方、日本語に関しては、言語研究を目的とする汎用的な統語解析情報をタグ付けしたコーパスは未だに存在しないため、コーパスにもとづく日本語の諸構造体の研究は英語などと比べて、はるかに立ち遅れている。日本語と世界の他言語とのコーパスにもとづく対照言語研究を行うためにも言語研究を目的とする汎用的なツリーバンクの開発は不可欠であり、その基礎となるアノテーションの研究が急務である。また、生きた日本語の膨大なデータの中から構造体や表現パターンにもとづく柔軟な検索が可能になることにより、外国人に対する日本語教育に関しても大きな便宜を与えることが出来る。

プロジェクトサブリーダーの吉本とプロジェクトフェローのバトラーは、平成 13 年度までの科学技術振興機構との共同研究を通じて、日本語文統語解析情報の意味処理によって、文統語・意味解析情報をタグ付けするコーパスの開発手法を編み出した (Butler 2015, Butler et al. 2016)。これによって、文の中核的な文法情報である語句間の依存関係 (dependency) を正確に抽出し、日本語文法研究にとって画期的な手段を提供することが出来る。これまでに言語理論や日本語学など種々の観点からなされてきた精緻な日本語文法研究に関する考察を大量データによって検証することが可能になろうとしている。これを受けて、国立国語研究所では本年度より、文統語解析情報に加えて意味解析情報をタグ付けした日本語コーパス NINJAL Parsed Corpus for Modern Japanese (略称 NPCMJ, 別称「けやきツリーバンク」) の構築を開始した。

2　アノテーションの原則

ツリーバンク開発に当っては、Annotation Manual for the Penn Historical Corpora and the PCEEC (Santorini 2010) の規約に従う。これは、Penn Treebank の解析規約を修正したもので、極力フラットな統語構造を採用してノードの数を減らすことと、名詞句、節等に必要に応じて機能情報 (主語、目的語、時間副詞句、節の様々な機能等) をタグ付けすることを特色としている。構造的曖昧性が問題になる場合の多くで統語的埋め込みをフラットなままに未指定とすることが出来るので記述しやすく、また有用な文法情報に富んでいる。さらに、多くの言語 (英語、フランス語、ポルトガル語、イディッシュ語等) のコーパス開発に採用されていることから、それらにおける多様な文法現象の取り扱いが日本語ツリーバンクの作成に当たって参考になり、さらに外国人研究者にも利用しやすいという利点が生じる。

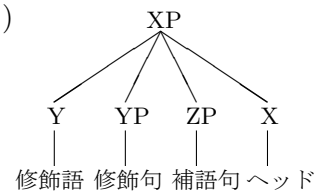
上記のペン通時コーパスの解析規約では、文の統語構造をラベル付きのカッコによって表示する。文のすべての単語に対して、品詞情報を表す語彙的ラベル (N, ADJ, VB, P など) がタグ付けされる。句 (phrase) に対しては、句レベルのラベルである NP, PP, IP 等が付加される。本コーパスの特色の 1 つである機能情報は、必要に応じて句レベルのラベルの後に NP-SBJ (主語名詞句)、IP-REL (関係節) のように付加される。

以下に、ペン通時コーパス方式のアノテーションの特色と利点および日本語への適用例を説明する。

(i) すべての種類の句が同一のフラットな構造を取る

(1) のスキーマに見られるように、句のヘッド (N, P, ADJ 等) がつねにそれと同一カテゴリーの句 (NP, PP, ADJP 等) を投射する。句のレベルとヘッドとの間に中間的なノードは存在せず、修飾語句 (modifiers) や補語句 (complement) とは同一レベルの姉妹となる。例えば、主文を表すノード IP-MAT は、動詞や述語を構成する他の要素、および副詞等を直接支配する。 さらに、日本語の場合、ヘッドはつねに句の右端にあらわれる。このようにすべての種類の句が同一のフラットな構造を取ることにより、木構造の検索や変換がきわめて簡単に行える。

(1)



また、これにより、異なるスコープ (作用域) 間の包含関係が見られる節の内部において、統語構造の埋め込みによる干渉を防ぐことが出来る。スコープの出現順位に従い、最初のものほど広いスコープを持つというディフォールトのスコープ包含関係を設定し、これに反するスコープ関係のみを記述することによって、柔軟なスコープ包含関係の指定を可能にする。

(ii) 句や節の機能タグ付けをすることにより、より正確な統語情報を提供することが出来る

これにより、構造の曖昧性を克服して、述語-項関係にとどまらないより複雑な構文により表現された意味情報を抽出することが可能になる。通常の関係節をともなう文 (2a) (主名詞「写真」は関係節の内部で目的語の働きをする) の統語木 (2b) に対して、埋め込まれた節が主名詞「写真」の内容を表している文 (3a) の統語木は (3b) のように表される。

(2) a.　昨日撮った写真がかかっていた。

b.
```
(IP-MAT (PP (NP-SBJ (IP-REL (NP-OB1 *T*)
                             (NP-SBJ *pro*)
                             (NP-TMP (N 昨日))
                             (VB とっ)
                             (AXD た))
                    (N 写真))
            (P-CASE が))
```

```
(VB かかっ)
(P て)
(VB2 い)
(AXD た)
(PU 。))
```

(3) a. 子供が泳いでいる写真がかかっていた。

b.
```
(IP-MAT (PP-SBJ (NP (IP-EMB (PP (NP (N 子供))
                                (P-CASE が))
                            (VB 泳い)
                            (P-CONJ で)
                            (VB2 いる))
                    (N 写真))
                    (P-CASE が))
(VB かかっ)
(P-CONJ て)
(VB2 い)
(AXD た)
(PU 。))
```

(2b) で通常の関係節を表すアノテーション IP-REL が使われる一方、(3b) では埋め込まれた節を表す IP-EMB が使われることによって、(2a) のような、いわゆる「内の関係」の関係節と、(3a) のような「外の関係」の関係節との違いを捉えることが可能となる。これについては、次節で詳しく述べる。

3 なぜツリーバンクか

日本語に関して従来利用可能なコーパスは、文を文節に分割した上で形態素情報をタグ付けし、文節間の係り受け関係を捉えようとするものが中心であった。しかし、文の理解、すなわち意味解析の手がかりとするには文節は不適切であり、句構造を採用しなければならないことは、すでに 1960 年代から日本語研究者たちによって主張されていることである。私たちの NPCMJ は十分な量の日本語テクストデータに対して句構造解析情報をタグ付けした最初の本格的なコーパスである。本節では、その長所について解説する。

NPCMJ のアノテーションにより、例えば、「内の関係」「外の関係」の 2 種類の関係節修飾や助詞「と」の異なる働き (引用/条件、他) を区別することが可能になる。上記 (2b) と (3b) の関係節と主名詞に相当する統語木を図 1a, b に示す。両者の違いは、トレース (*T*) の有無および関係節に与えられるラベル IP-REL と IP-EMB との違いとして示されている。ここで、IP-REL は主名詞が関係節において格役割を果たす関係節を示し、IP-EMB は主名詞が関係節中で格役割を果たさず、後者が前者の内容を表すことを示している。これらにもとづいて、意味解析処理では図 1a のアノテーションは

$\ldots$ 写真 $(x) \wedge$ とる $(e_1, pro, x) \ldots$

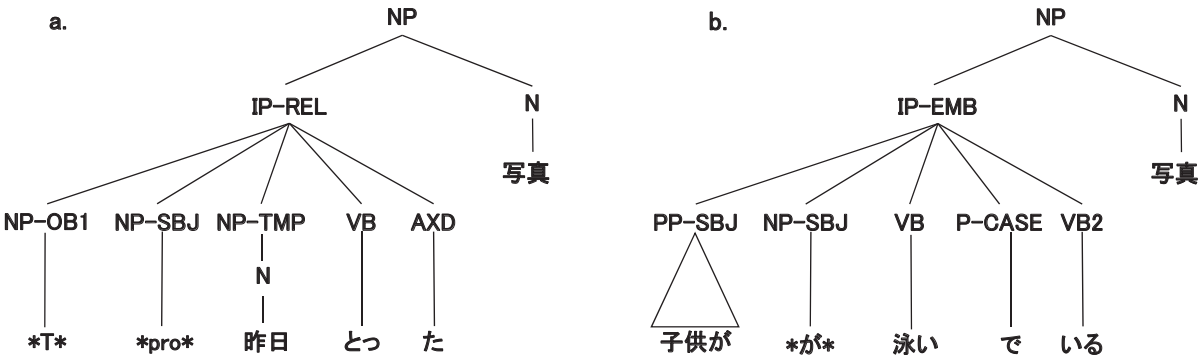
の論理表示を出力する。ここで、述語「とる」の目的項として、「写真」を表す x が入っている。他方、図 1b のアノテーションの出力は

$\ldots$ 子供 $(x) \wedge$ 写真 $(y,$ 泳いで_いる $(e_1, x)) \ldots$

であり、「子供が泳いでいる」という節の意味は述語「写真」の項として埋め込まれている。

従来の文節にもとづくコーパスにおいては、両方の例ともに [動詞 + 助動詞] [名詞] のように分析されるだけであり、両者の違いを捉えることは出来ない。

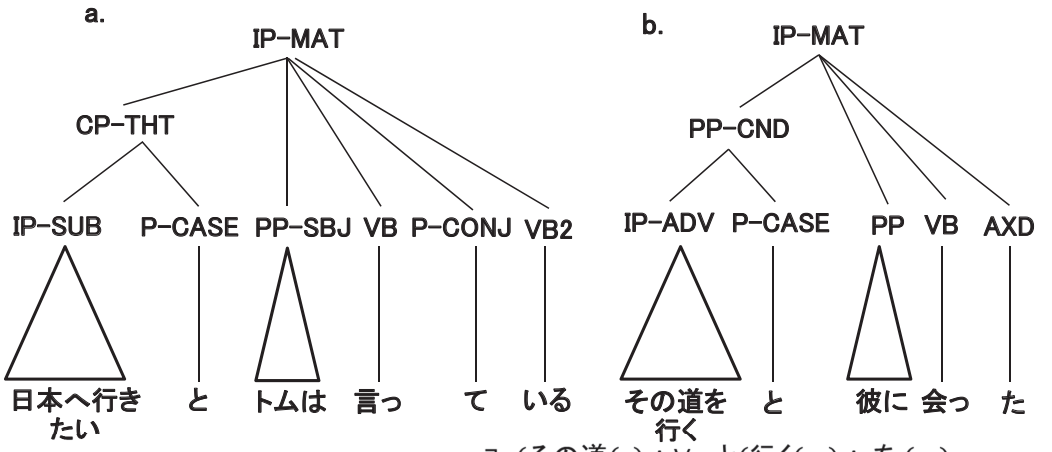
また、接続助詞「と」が導く節には引用節と条件節とがあるが、これらについても従来の文節コーパスは [動詞句 + 助詞] のように分析するだけで、意味上の区別には役立たない。両者の曖昧性解消のために NPCMJ で行っているアノテーションを図 2a, b に示す。図 2a では当該の節は CP-THT のカテゴ



$\cdots$写真(x) ∧ とる $(e_1, pro, x)\cdots$

$\cdots$子供(x) ∧ 写真 $(y,$泳いで_いる$(e_1, x))\cdots$

図 1: 内の関係と外の関係



∃x(その道(x) ∧ ∀e_1と(行く(e_1) ∧ を $(e_1) = x$,

∃e_1言っている$(e_1$,トム,∃ e_2(行きたい (e_2) ∃$e_2 y$(彼 : $y = ?$ ∧ 会っ (e_2, y) ∧ に $(e_2) = y$)))

∧ へ $(e_2) =$ 日本))

図 2: 助詞「と」の曖昧性

リーによって引用節であることが示されているのに対して、図 2b では PP とされていることに加え、曖昧性解消のために *CND* が付加されることにより、前者と異なる条件節であることが示されている。

4　文法研究との関係

アノテーションに際しては、これまでに行われてきた、日本語の多様な構文研究の成果を利用する。しかし、アノテーションとは、それぞれの事象について既存の解析法の 1 つを選んで当てはめれば行えるというものではない。

第一に、これまでの構文研究は特定の研究者にとって関心のある一部の構文を取り上げて論じたものが多い。これに対して、本プロジェクトにおけるようなコーパスのアノテーションは、大規模なデータをすべて扱う必要がある。特定の構文を説明するには有効な規則でも、そのままでは他の構文の解析と矛盾を起こすことがある。このような時には、両方を説明できる別の規則を立てるか、あるいは正確さに劣る規則で我慢しなければならないこともある。また、言語データの中には、研究者がこれまで無意識にかあるいは意識的に避けてきた、規則で取り扱いにくいものも多数含まれている。

さらに、アノテーションはデータ全体を通じて一貫したものでなければならない。そのためには、できるだけ明確なアノテーション基準を作る必要がある。機能語 (助詞、助動詞) や構文の分類に関して、それが意味や文脈にもとづくデリケートなものである場合には、このような観点から比較的大雑把な分類を採用しなければならないこともある。

また、ある 2 つのカテゴリーへの分類が一般に行われている場合、分類自体は有効でも、両者の境界が判然としない場合もある。先に取り上げた、関係節の内の関係 (IP-REL) と外の関係 (IP-EMB) への区別は、実はそのような事象の一例である。以下の文で、

(2)　a.　火事が広がった 原因 は空気が乾燥していたことだ。(寺村 1975)

　　b.　私が 16 だった とき、彼女はまだ 7 つでした。(益岡・田窪 1992)

(1a) の「原因」が導く関係節は外の関係にあるとされることが多いが、この節は「その 原因で 家事が広がった」と言い換えられることから、内の関係と見ることもできる。同様に、(1b) のように時間関係を表す「とき」節は、高度に抽象的な統語情報 (いわゆる「相対時間」等) に関わることから外の関係としたいところだが、内の関係との見方も可能である。

以上のようなことから、アノテーション作業におけるカテゴリーの決定は、分かりやすくはあっても、必ずしも言語学的に有意味でも正確でもないかも知れない「ヒューリスティクス」を用いて行う必要が生じる。しかしそのことは必ずしも、そのようにして行われる分類やアノテーションの言語学上の価値を下げるものではないと考える。第一に、このような分類方法が必要だということは、従来の言語学者による分類が不完全なものであったということを教えてくれる。第二に、表層的だという傾向があるとはいえ、ともかくそのような一貫した分類を行うことにより、これまでは不可能であった、膨大な言語データを対象とする網羅的な文法研究が可能になるからである。

5　おわりに

意味解析のための適合性、記述の客観性と一貫性、言語使用の実態の反映および検索利用の容易性、という異なる要因のバランスを配慮した開発を行う。形態素解析は、主として、名詞については BCCWJ の長単位、述語句については短単位にもとづいて行い、述語句については形態素ごとの検索がほぼ可能である。言語データとしては、著作権上の問題の生じない Wikipedia 記事や問題が解決された新聞記事等を取り上げる。

6 年間のプロジェクト終了後には、約 5〜6 万文の日本語テクストについて、統語・意味解析情報付きコーパスを完成し、国立国語研究所のウェブサイト

`http://npcmj.ninjal.ac.jp/`

から公開する予定である。今年度の成果として、近日中に約 1 万文のコーパスを上記サイトより公開する。その内訳は以下のとおりである。

出典	例文数
『河北新報』記事	4,243
Wikipedia 記事	2,752
新約・旧約聖書	1,659
益岡・田窪 (1992) 例文	1,378
合計	10,032

例文のローマ字化も行い、また初心者も利用できる簡便なインタフェースとともに公開する予定である。これにより、コンピュータ処理の習熟度や言語の壁を越えた幅広い利用が期待できる。

参考文献

赤瀬川史朗・プラシャント・パルデシ・今井新悟 (2016)『日本語コーパス活用入門 — NINJAL-LWP 実践ガイド』大修館書店。

Butler, Alastair , 吉本啓, 岸本秀樹, プラシャント・パルデシ (2016)「統語・意味解析情報付き日本語コーパスのアノテーション」『言語処理学会第 22 回年次大会発表論文集』, pp.589-592, 東北大学。

Butler, Alastair. (2015) *Linguistic Expressions and Semantic Processing: A Practical Approach.* Springer.

Santorini, B. (2010) Annotation Manual for the Penn Historical Corpora and the PCEEC (Release 2). Tech. rep., Dep. of Computer and Information Science, University of Pennsylvania.

寺村秀夫 (1975)「連体修飾のシンタクスと意味—その 1—」『日本語・日本文化』4 号、大阪外国語大学留学生別科; 寺村秀夫 (1992)『寺村秀夫論文集 I —日本語文法編—』くろしお出版。

益岡隆志・田窪行則 (1992)『基礎日本語文法—改訂版—』 くろしお出版。