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

アラステア・バトラー　吉本 啓　岸本 秀樹　プラシャント・パルデシ
東北大学　東北大学　神戸大学　国立国語研究所
ajb129@hotmail.com

1 はじめに

日本語に関して現在のところ公開されているコーパスは，文を文節に分解して形態論情報をタグ付けしたものや，せいぜい文節間の係り受け関係を示したものにとどまっている．人文系・工学系の区別を問わず，世界の趨勢に歩調を合わせて高度の日本語研究を行っていくには，日本語の文の統語解析情報をタグ付けしたコーパス (ツリーバンク) の開発が急務である．国立国語研究所では，平成 28 年 4 月より，現代日本語の書き言葉テクストに対し文の統語・意味解析情報をアノテートした NINJAL Parsed Corpus for Modern Japanese (NPCMJ) の本格的な構築を開始する．本稿では，まず本コーパスプロジェクトの概要について説明し，文法研究を目的とするテクストデータのピンポイントの検索および意味解析の抽出という主要目標を達成するためのアノテーション方式について述べる．

2 プロジェクトの概要

国立国語研究所の第 3 期共同研究プロジェクト「統語・意味解析コーパスの開発と言語学研究」(平成 28 年 4 月～平成 34 年 3 月) では，現代日本語書き言葉テクストを対象として統語解析木と論理意味表示 (述語論理式) をアノテートしたコーパス NPCMJ の本格的な構築を本年 4 月より開始する．

現在世界の主要言語について Penn Treebank (Bies, et al. 1995) 方式のツリーバンク (統語解析情報付きコーパス) が作られ，言語学および言語処理の研究に目覚ましい成果を挙げている．しかし日本語については十分な規模の公開されたツリーバンクは存在しない．NPCMJ はこのような日本語研究の遅れを挽回し，多様な日本語の機能語，句，節および複雑な構文を大量の言語データから検索・抽出して研究することを可能にすることを目的としている．成果として得られるコーパスは，言語処理の技術を持たない人でも簡単に利用できるインタフェースとともに，国立国語研究所のホームページから一般公開する予定である．

言語データとしては，新聞記事をはじめとする，構文のしっかりした現代の書き言葉を採用する．すでに東北地方のブロック紙である『河北新報』を発行する河北新報社より同紙記事公開の許可を得ており，NPCMJ のテクストデータの一部として利用する予定である．

また，汎用性を優先させるため，特定の形式言語理論にコミットせず，中立的なアノテーションのスキーマを採用する．具体的には，本プロジェクトの目的に合致する，Penn Treebank の一つのバリエーションであるペン通時コーパス (Penn Historical Corpus; Santorini 2010) のアノテーション方針に従う．

3 構築方法

言語テクストはまず，MeCab, UNIDIC および Comainu を使用して形態素解析に掛けられる．論理意味表示を行うという目的からは，助詞や助動詞は独立した単語として扱うのが適切である．他方，複合動詞や複合名詞等の複合表現は重要な問題を提起しているものの今は立ち入らないことにし，1 つの単語として扱うことにする．国語研究所で確立されたセグメンテーションの単位のうち，短単位 (Maekawa et al. 2014) として助動詞や助動詞をタグ付けし，複合表現については長単位に従うという，2 つの基準を混在させたセグメンテーションを採用する．

形態素解析の結果は，Bitpar Probabilistic Context-Free Grammar にもとづいて作られた統語解析器に掛けられる．その解析結果は人手により誤りを修正され，その結果が文統語解析情報としてコーパスに利用される．修正結果はまた，統語解析プログラムにフィードバックされる．

上記の統語解析情報は，意味解析システムへの入力としても利用され，これから文の論理意味表示 (イベント論理にもとづく述語論理式) が自動的に生成され，

文意味解析情報としてタグ付けされる．文の意味解析 (文の評価) は，バトラーの提唱するスコープ制御理論 (Scope Control Theory; Butler 2010, Butler 2015) をインプリメントした意味解析システムによって行われる．同理論は，文の適切性が意味論的な条件によって決定されるという考えを基礎としている．

例文 (1a) の統語解析情報を (1b) に，またその意味解析情報を (1c) に示す．

(1) a. となりに座った人と話しました．

b.
```
(IP-MAT (NP-SBJ *speaker*)
        (PP (NP (IP-REL (NP-SBJ *T*)
                        (PP (NP (N となり))
                            (P に))
                        (VB 座っ)
                        (AXD た))
                (N 人))
            (P と))
        (VB 話し)
        (AX まし)
        (AXD た)
        (PU . ))
```

c. $\exists x_1 e_5 e_6 x_4 x_2 ($
$x_1 = \text{speaker} \wedge$ となり$(x_4) \wedge$
座っ_た$(e_5, x_2) \wedge$ に$(e_5) = x_4 \wedge$ 人$(x_2) \wedge \text{past}(e_5) \wedge \text{past}(e_6) \wedge$ 話し_まし_た$(e_6, x_1) \wedge$ と$(e_6) = x_2)$

(1b) に見られるように，本方式の統語アノテーションにおいては，関係節に欠如している主語名詞句が関係節の被修飾句である「人」であることは，形式文法において通常同一性を示す手段であるインデクスなどの手段で示されているわけではない．

ここで示されている統語情報は，関係節と主名詞の修飾 - 被修飾関係と関係節内部で主語名詞句が欠けている (トレースとなっている) ということだけである．この「表層的」な統語情報にもとづき，意味解析によって (1c) の述語論理式が得られる．述語論理式において，述語「座る」と「人」とは変項 x_2 を共有しているが，このことは「座る」の主語が「人」であることを表している．すなわち，統語解析情報はインデクスを持たず表層的であっても，意味解析によって同一性に関する情報が補われるのである．このことは埋め込み文の関係節化などのより複雑な例についても行われるし，また非有界依存構文 (unbounded dependency) 一般についても同じ方法が適用される．このようにして，統語アノテーションにおいてインデクス付けを省略し，現実的な作業量の範囲内でコーパス構築を行うにもかかわらず，複雑な構文についても一貫して依存関係に関する情報を提供することが可能になる．

4 タグ

表 1 に，記号を除く品詞タグの一覧を掲げる．全部で 26 種類あるこれらのタグは単語に対して与えられるタグであり，概ね常識的なものだが，助動詞類が AXD (過去助動詞「た」)，MD (モーダル助動詞)，NEG (否定辞)，PASS (受動助動詞)，VB2 (補助動詞)，および AX (他の助動詞) に細分化されていること，また動詞についても VB (動詞語幹) と VBO (軽動詞) とが区別されていることを特色とする．

表 2 に，統語タグ (句や節のカテゴリーを示すもの) を構成する 13 種類のタグのリストを掲げる．これらのうち，NP に対しては任意的に，ハイフンに続けて機能を表示するラベルを付加することができる．CP および IP については機能ラベルの付加が必須である．

ADJI	い-形容詞	NPR	固有名詞
ADJN	な-形容詞	NUMCL	助数詞
ADJT	たる-形容詞	P	助詞
ADV	副詞	PASS	受動助動詞
AX	助動詞	PRO	代名詞
AXD	過去テンス標識	Q	量化詞
CARD	数詞	VB	動詞 (語幹)
CONJ	並列接続詞	VBO	軽動詞
D	限定詞	VB2	補助動詞
FW	外来語	WADV	疑問副詞
INTJ	間投詞	WCARD	疑問数詞
MD	モーダル助動詞	WD	疑問限定詞
N	名詞	WPRO	疑問代名詞
NEG	否定辞		

表 1: 品詞タグ

ADJP	形容詞句	NML	中間名詞節
ADVP	副詞句	NP	名詞句
CONJP	接続詞句	NUMCLP	助数詞句
CP	従属節	PP	後置詞句
FRAG	断片	PRN	カッコ挿入句
INTJP	間投詞句	QP	量化詞句
IP	節		

表 2: 統語タグ

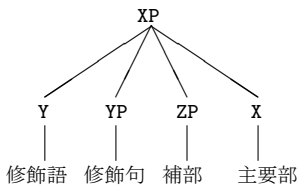
例えば，NP-SBJ および NP-OB1 はそれぞれ主語名詞句および直接目的語名詞句を示す．CP-THT は引用節を，また IP-REL は関係節を意味する．機能タグは，統語解析情報の曖昧性を克服し，正確な意味解析情報を抽出するための手がかりとして利用される．

5　一般的な原則

統語解析木において，節の構造は一般にフラットであり，複数の枝分かれノードを持つ．その中で，IP (主節) がすべての動詞や助動詞およびその他の文レベルの構成素を直接支配する．このように平坦な統語構造を採用するのは，句のレベルと主要部の間の中間的なノードがスコープに干渉すること防ぎ，また柔軟なスコープ包含関係を可能にするためである．しかし，必要な場合には，NPCMJ の解析木を自動的に二分木に変換することは容易に行うことができる．

(2) に，本プロジェクトで用いる一般的な句の内部構造を示す．具体的には，図の 'X' の部分に N, P, ADJI 等が入り，主要部 (ヘッド) と句ラベルの組み合わせが N-NP, P-PP, ADJI-ADJP 等になる．ただし，主要部が動詞 (VB) の場合，句 (節) 全体に与えられるラベルは IP となる．句の種類を問わず類似の構造を持つという考えは X バー理論に従っているが，通常の同理論と異なって，中間レベルの構造 (N' や VP) は決して明示されない．このようにフラットな統語構造を採用しても，前節で述べた機能表示によって修飾部と補部との区別を行うことができる．

(2)



6　アノテーションの特徴

主要目標の 1 つである意味解析情報の抽出のために，一部で独自のアノテーション方式を導入している。

6.1　格情報

日本語の主語，直接目的語，間接目的語は「は」「が」「を」「に」等の助詞により表示されるが，意味役割と助詞とは多対多に対応する．そのため，これらの 3 つの格については，当該の後置詞句の直後に (SBJ *)，(NP-OB1 *を*) のようなノードによって意味役割を表示し，曖昧性を解消する．

```
(3)  (IP-MAT (PP (NP (NPR 花子))
                 (P は))
             (NP-SBJ *)
             (PP (NP (N 長女))
                 (P を))
             (NP-OB1 *を*)
             (IP-INF (PP (NP (N 買い物))
                         (P に))
                     (VB 行か))
             (VB2 せ)
             (AXD た)
             (PU . ))
```

6.2　従属節

従属節の内で，等位節と条件節とを区別する．これは，等位節が連言 (∧) によって主節と結合されるのに対し，条件節は含意 (→) によって結合される上に全称量化を受け，意味表示としては大きく異なるからである．前者では従属節の後に (CRD *) を，後者では (CND *) をタグ付けして区別する．(4) を参照のこと．

6.3　空要素

日本語においては，主語や目的語が省略されることが非常に多い．このような「ゼロ代名詞」のタグ付けは，文の統語構造の確定および単文の意味の正確な表示という両方の点から不可欠である．空要素一般について，本アノテーション方式では，統語論的・意味論的性質にもとづいて，以下のように分類する．

(i) 関係節のトレース標識
(ii) 虚辞 (expletive)
(iii) ゼロ代名詞
(iv) 不連続構造の標示のためのトレース

このうち，(i) については 3 節で述べた．

(ii) は，天候を表す動詞などが従属節として埋め込まれる場合，意味的な主語が存在せず，したがって主節中の項によるコントロール関係を排除したい場合に用いられる．

(iii) は，必須格の省略のうち，照応詞 (アナフォラ) として用いられるものである．その指示対象の違いによって，以下のように異なるタグ付けがなされる．

arb　一般的非人称指示

pro　定の指示 (small pro)
hearer　聞き手を指示
speaker　話し手を指示
speaker+hearer　話し手および聞き手を指示
speaker+pro　話し手および定の個体を指示

pro は談話文脈中の個体を指示するか，あるいは同一の文中の要素を先行詞として取る．*hearer*, *speaker*, *speaker+hearer*, および *speaker+pro* は *pro* の特殊な場合であり，条件が満たされる場合，優先して使用される．

(iv) は，不連続構造を表示する必要から，例外的にインデクス付けを行う空要素である．これは，句のレベルを横断する外置，スクランブリング，および他の転置に関して使用される．転置元に *ICH* およびインデクスをタグ付けし，また転置先の構成素のラベルにも同じインデクスを付加して，転置された構成素がどの位置で解釈されるかを示す．

しかし，必須格が明示的に表現されない場合のすべてがゼロ代名詞として扱われるわけではない．従属節内部の必須格が表現されていない場合，従属節の種類によって，主節の項をディフォールトとして継承する．このように考えるのは，必須格の「省略」が見られるたびにゼロ代名詞をタグ付けしていたのではその指示対象をいかに同定するかという問題が生じ，複文の階層にもとづいて非明示要素のディフォールト的解釈を行うという日本語の実情 (南 1974) に合わなくなるためである．

不定詞節，副詞節，テ-従属節，および名詞化節と形式名詞埋め込み節については，それらの節の中の主語の指示対象に関する，主節の必須格項をコントロール元とする継承がディフォールトとして行われる．すなわち，これらの従属節の主語が明記されていない場合，基本的に主節中の間接目的語，直接目的語，主語の優先順で指示対象が同一であるとしてディフォールト的意味解釈が行われる．ただしその際，これらの主節の項と従属節との相対的順位や従属節の種類に応じて異なる条件が課せられる．(4) は，条件節が主節の主語 (聞き手) を継承する例である。

(4)
```
(IP-IMP (NP-SBJ *hearer*)
        (PP (IP-ADV (VB 行か)
                    (NEG ない))
            (P にしろ))
        (CND *)
        (PU , )
        (NP-OB2 *speaker*)
        (ADVP (ADV 後で))
        (PP (NP (N 電話))
            (P を))
        (NP-OB1 *を*)
        (VB 下さい)
        (PU . ))
```

7　結論

国立国語研究所における，現代日本語書き言葉テクストに対する統語・意味解析情報付きコーパスの開発計画について説明した。平成 34 年 3 月までに 5 万文以上のアノテーションを完成し公開する予定である．

この研究は国立国語研究所フィージビリティスタディ型共同研究プロジェクト「日本語テクストのツリーバンクアノテーション法の開発」(平成 26〜27 年) の支援を得て行われた．また，東北大学国際文化研究科附属言語脳認知総合科学研究センターの援助を得た．予備研究は科学技術振興機構戦略的創造研究推進事業さきがけ「自然言語テクストの高精度で頑強な意味解析とその応用」(平成 22〜25 年) で行った．

参考文献

Bies, A., et al. Bracketing guidelines for Treebank II style Penn Treebank project. Univ. Pennsylvania Computer and Information Science Dept. 1995.

Butler, A. *The Semantics of Grammatical Dependencies.* Emerald. 2010.

Butler, A. *Linguistic Expressions and Semantic Processing.* Springer. 2015.

Kurohashi, S. and M. Nagao. Building a Japanese parsed corpus - while improving the parsing system, A. Abeille, ed., *Treebanks: Building and Using Parsed Corpora.* Kluwer. 2003.

Maekawa, K, et al. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation* 48(2). 2014.

南不二男『現代日本語の構造』大修館．1974.

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