日本語ツリーバンクのアノテーション方針

吉本 啓†‡　周 振‡　小菅 智也*　大友 瑠璃子†　Alastair Butler♠†
†東北大学高等教育開発推進センター　‡東北大学大学院国際文化研究科
*東北大学大学院情報科学研究科　♠科学技術振興機構戦略的創造研究推進事業さきがけ
kei@compling.jp

要旨

日本語文に論理意味表示をタグ付けしたコーパス開発のために必要な、統辞解析情報のアノテーションの原則について解説する。ペン通時コーパスの解析規約に従って、フラットな文構造を採用し、またラベルの一部に機能情報を追加する。さらに、意味解析への入力の必要から、機能語を1語として扱ったり、ゼロ代名詞をタグ付けする等の特色を持っている。

1　はじめに

日本語の無制約のテクストに対して、述語論理式を付加した意味表示コーパス欅ツリーバンク (Keyaki Treebank) を構築している (Butler et al. 2012a, Butler et al. 2013)。そのために前提として必要な、日本語の統辞解析情報付きコーパス（ツリーバンク）を作成中である。各文の統辞解析情報が得られれば、スコープ制御理論　(Scope Control Theory (SCT); Butler 2010) を実装したシステムに入力することによって、自動的に論理意味表示 (述語論理式) を得ることが出来る。最終的に、約4万文について統辞・意味情報をタギングする予定だが、このような大きさの意味コーパスは、日本語はもちろんのことどんな言語についても存在せず、完成すれば、これまでにない深さと広さを兼ね備えた言語研究に貢献することが出来る。本発表ではその前提としての、日本語統辞解析情報アノテーションのために立てられた方針について説明する。

2　これまでの研究

日本語の文に対して統辞解析情報を付加したコーパスはいくつかあるが、その代表である京都大学テキストコーパス (Kurohashi et al. 2003) をはじめとして、その多くは文節にもとづくものであり、筆者らが目指す文の意味の自動解析に利用するには問題がある。京都大学テキストコーパス第4版のうち5,000の文については、格、照応および指示情報が付加されており、これにもとづいて格フレーム（述語-項関係）に関する情報を読み取ることができる。さらに筆者らは、単純な述語-項関係を超えて、文が実際に表示する複雑な統辞情報を京都大学テキストコーパスにもとづいて得ようと試みた (Butler et al. 2012b)。しかし、文節表記の壁は厚く、個々の文節を超えてシステマティックに情報を補充する方法を見出すことは出来なかった。また、わずかな変更が文節間の依存関係全体に波及効果を与えることも多い。そこで筆者らは、日本語ツリーバンクをゼロから構築することに方針転換した。

3　ペン通時コーパスのアノテーション方式

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、極力フラットな統辞構造を採用してノードの数を減らすことと、名詞句、動詞句、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辞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、文の統辞構造をラベル付きのカッコによって表示する。文のすべ

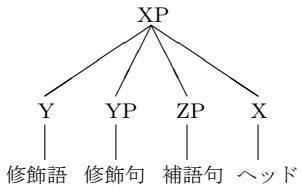
ての単語に対して、品詞情報を表す語彙的ラベル (N, ADJ, VB, P など) がタグ付けされる。句 (phrase) に対しては、句レベルのラベルである NP, PP, IP 等が付加される。本コーパスの特色の 1 つである機能情報は、必要に応じて句レベルのラベルの後に NP-SUB (従属節)、IP-REL (関係節) のように付加される。

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われる。

(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、統辞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することにより、より正確な統辞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) の統辞木 (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) の統辞木は (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 (IP-REL (NP-SBJ *T*)
                        (NP-TMP (N 昨日))
                        (VB とっ)
                        (AXD た))
                (N 写真))
        (P が))
(NP-SBJ *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 (NP (IP-EMB (PP (NP (N 子供))
                            (P が))
                        (NP-SBJ *が*)
                        (VB 泳い)
                        (P で)
                        (VB2 いる))
                (N 写真))
        (P が))
(NP-SBJ *が*)
(VB かかっ)
(P て)
(VB2 い)
(AXD た)
(PU 。))
```

(2b) と (3b) を SCT システムに入力して自動意味評価することにより、文 (2a) と (3a) の意味表示は、それぞれ (2c) と (3c) のように与えられる。

(2) c. $\exists x t_1 e_1 e_2$(
写真 (x) ∧ とっ(e_1, x) ∧ 時間$(e_1) = t_1$ ∧ 昨日 (t_1) ∧ past(e_1) ∧ past(e_2) ∧ かかっ_い (e_2, x))

(3) c. $\exists x y e_1 e_2$(
子供 (x) ∧ 写真 $(y,$ 泳い_いる $(e_1, x))$ ∧ past(e_2) ∧ かかっ_い (e_2, y))

関係節と主節との意味的関係は、(2c) では ∧ で連結される並列関係である一方、(3c) では関係節の意味の主名詞の意味への埋め込み関係となり、本質的に異なる。このような差を正しく捉えることは、(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可能となる。

4 アノテーションの原則と具体例

4.1 ラベル

現在構築中の日本語ツリーバンクでは、タグ付け基準の客観性・一貫性、日本語使用の実情、および意味表示からの要請、という時には互いに矛盾も生じるそれぞれの条件を最大限満たし、バランスの取れたコーパス開発の方針の確立を目指している。

語彙ラベル (品詞情報) のタギングは、上記のようにペン通時コーパスのものを基本とし、益岡・田窪 (1992) を参考にしつつ、意味情報付きコーパスの開

発という独自の目的に適うように行っている。MeCabを使って自動形態素解析した結果を人手で修正している。語彙ラベルとしては、以下の 23 種類 (記号を除く) を使用する。

(4) ADJ (形容詞), ADV (副詞), AX（助動詞), AXD（助動詞「た」), CARD (数詞), CONJ (接続詞), D (限定詞), FW (外国語), H (接頭辞), INTJ (間投詞), MD (モーダル助動詞), N (普通名詞), NEG (否定辞), NPR（固有名詞), NUMCL (助数詞), P (助詞), PRO (代名詞), Q (数量詞), VB (動詞), VB0（軽動詞), VB2（補助動詞), WD (疑問限定詞), WPRO (疑問代名詞), WADV (疑問副詞)

句に対しては、次の 12 種類のラベルが付加される。

(5) ADJP (形容詞句), ADVP (副詞句), CONJP (並列句), CP (節), FRAG (断片文), IP (節), NP (名詞句), NUMCLP (助数詞句), NX (名詞、名詞句のいずれとも決めがたいもの), PP (後置詞句), QP (数量詞句), VP (動詞句)

このうち、CP, NP および IP は機能を表すラベルとともに用いることが出来る。CP と IP は機能ラベルが必須である。CP は疑問節 CP-QUE となるか、または CP-THT の形で埋め込まれた名詞節として用いられる。NP は ADT (付加語)、ADV (副詞)、DIR (方向格)、LOC (場所格)、MSR (度量句)、OB1（直接目的語)、OB2（間接目的語)、SBJ (主語)、SBJ2 (2 つ目の主語)、POS (所有格)、SUM (集合)、TMP (時間格) の 12 種類の機能タグを付加して使用される。IP の機能タグは、ADV (副詞節) 、EMB (埋め込み節) 、IHR (ヘッド内在関係節)、MAT (主文) 、REL (通常の関係節) 、SUB (従属節) および TE (接続助詞「て」により導入される従属節) のいずれかである。

4.2 重要な原則

欅ツリーバンクのアノテーション方針のうち、独特かつ重要なものについて説明する。

(i) いくつかの単語が緊密に連結して 1 つの機能語として働くものは、1 つの助詞 (P) として扱う。

これは、最終目的である文意味解析の便宜のためである。この中には、「として」「について」「に対して/対する」「に関して/関する」等が含まれる。このうち、「として」「について」には助詞プラス動詞テ形としての用法もあり、構造的にあいまいなものとして取り扱う。また、通常形式名詞とされる「ため」「おかげ」「せい」「あまり」についても、「のために」のように 1 つの機能語に相当する用例については、1 語の P とする。

(ii) いくつかの単語が緊密に連結し 1 つのモーダル的機能を果たすものは、1 つの助動詞 (MD) とする。

これも、文意味解析の便宜のために行う。例えば、「なければならない」は 1 個のモーダル助動詞 MD とする。「開こ うとする」の下線部分は MD-AX+P+VB とタグ付けするが、これは、全体で 1 つの MD であり、それが助動詞、助詞および動詞から構成されることを示す。「解決する はずだ」の下線部分は MD+AX-LOW となる。これは、MD である「はず」の後に助動詞「だ」が後続していることに加えて、モーダル助動詞が狭いスコープを持つことを示している。

(iii) 後置詞句 (PP) が文中で主語や目的語として機能する場合、その直後に NP-SBJ, NP-OB1 または NP-OB2 を付加して、その文法機能を明示する。

これには、係助詞「は」や副助詞が付加されて格が明示されない場合を含む。しかし、格助詞「が」、「を」や「に」を伴う場合でも、格助詞により表示される文法役割があいまいなため、この方法によって格情報を明示する。会話文などでこれらの格助詞が省略された名詞句についても同様の表示を行う。例は (2b) の 7 行目および (3b) の 9 行目に示されている。

(iv) 関係節が修飾する名詞句において、主名詞が関係節の中で文法役割を果たす場合は、関係節内に空所に相当するノードを与えて文法役割を明示する。

例 (2b) を参照のこと。ここでは、1 行目の (NP-SBJ *T*) として示されている。

(v) 主語または目的語が動詞の必須格として求められるにもかかわらず文中で表現されていない場合の多くについて、それらをゼロ代名詞として明示する。

ゼロ代名詞のタギングを行うのは、依存関係の表示および述語-項関係の再構成に必要だからである。無主語文の主語のゼロ代名詞表示は必要無い。以下の (6a) の最初の行に、目的語をゼロ代名詞として表示した例を示す。(6b) の意味表示では、文脈中の先行詞と同定されるべき値として扱われている。

(6) a.
```
(IP-MAT (NP-OB1 *pro*)
        (PP (PP (NP (WPRO 誰))
                (P か))
            (P が))
        (NP-SBJ *が*)
        (VB 助ける)
        (MD だろう)
        (PU 。))
```

b. $\exists yx(\text{pro}{:}y \;=\; ? \;\wedge\; \text{だろう}(\exists e_1\, \text{助ける}(e_1, x, y)))$

主語や目的語が明示されなていなくても、それと同一指示の名詞句が文中に存在してコントロール関係にある場合は SCT のスコープ操作によってコントロール関係が意味論のレベルで補完されるため、ゼロ代名詞としてのタギングは行わない。以下に例を示す。

(7) a.
```
(IP-MAT (NP-SBJ *pro*)
        (PP (IP-TE (ADVP (ADV よく))
                   (VB 考え)
                   (P て))
            (P から))
        (PU 、)
        (VB ご返事)
        (VBO し)
        (AX ます)
        (PU 。))
```

b. $\exists x e_1 p_1$(
pro:x ＝ ? ∧
fact(p_1, $\exists e_2$(考え(e_2, x) ∧ よく(e_2))) ∧
ご返事_し_ます(e_1, x) ∧ から(e_1) ＝ p_1)

(vi) 例外的な場合を除き、インデックスは使用しない。

長距離依存のような複雑な構文でもインデックスが不要であることは SCT の大きな特徴であり、これによって、本研究で目指しているツリーバンク構築の作業量が著しく軽減される。また、表層的な自動統辞解析結果を SCT システムへの入力とし、述語論理式出力までのすべての過程を自動化することも視野に入れることが出来る。例外となるのは、外置 (extraposition)、数量詞遊離 (floating quantifier)、主要部内在型関係節のいずれかの構文で、意味処理上の要請から、語句をその実際の位置以外の場所と関係づける必要が生じる場合である。以下に、数量詞遊離の例を示す。(8a) で「それぞれ」がインデックスにより「隊員たち」を修飾する位置と関連付けされ、そのことにより (8b) で正しい意味表示を得ている。

(8) a.
```
(IP-MAT (PP (NP (Q *ICH*-1)
                (N 隊員たち))
            (P は))
        (NP-SBJ *は*)
        (Q-1 それぞれ)
        (PP (NP (PP (NP (PRO 自分))
                    (P の))
                (N 部署))
            (P に))
        (VB 帰っ)
        (P て)
        (VB2 行っ)
        (AXD た)
        (PU 。))
```

b. それぞれ x(隊員たち(x),
$\exists z y e_1$(
自分:z ＝ choose1(y) ∧
の_部署(y, z) ∧
past(e_1) ∧
帰っ_行っ(e_1, x) ∧ に(e_1) ＝ y))

5 結論

論理意味表示をタグ付けしたコーパス構築のための文解析情報アノテーションの基本的な原則について説明した。現在構築中のツリーバンクを順次以下のサイトで公開する予定である。

http://www.compling.jp/keyaki/

引用文献

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

Butler, A., et al. (2012a) “Keyaki Treebank: Phrase Structure with Functional Information for Japanese”, テキストアノテーションワークショップ, 国立情報学研究所

Butler, A., et al. (2012b) “Problems for Successful Bunsetsu based Parsing and Some Solutions”, 『言語処理学会第 18 回年次大会発表論文集』

Butler, A., et al. (2013) “Treebank Annotation for Formal Semantics Research”, *Proceedings of the Ninth International Workshop on Logic and Engineering of Natural Language Semantics*, pp. 210-222. JSAI International Symposia on AI.

Kurohashi, S. and M. Nagao. (2003) “Building a Japanese parsed corpus – while improving the parsing system”, A. Abeillé, ed., *Treebanks: Building and Using Parsed Corpora*, chap. 14. Kluwer Academic Publishers.

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

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.