統語・意味解析情報付き日本語コーパスの開発

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

1　はじめに

筆者たちが計画している統語・意味解析情報タグ付き日本語コーパスは，各文に対して句構造および述語論理による論理意味表示をアノテートしたコーパスである．その最大の特徴は，文中の語句間の完全な統語的・意味的リンク付けを行うことにより，文法情報の根幹をなす依存関係 (dependency) の抽出を可能にすることにある．これにより，様々な文法的条件を満たす構造体 (constructions; 統語構造と意味とのペア) の検索が可能になる．コーパスは，プログラミングや自然言語処理の技術を持たない言語研究者でも簡単に利用できるインタフェースとともに公開する予定である．以下では，研究の目標，コーパス開発の動機，構築方法，アノテーションの方式，特色と意義，および予定される成果と将来の展望について説明する．

2　研究の目標

現在のところ日本語に関するコーパスで公開されているものは，文を文節に分解した上で形態論情報をタグ付けしたものがほとんどである．構文に関するアノテーションとしては，せいぜい文節間の係り受け関係を示したものしか見当たらず，文法研究に利用するには限界がある (Kurohashi and Nagao 2003, Maekawa et al. 2014)．私たちは，様々な形で文法情報を必要とする研究者が機能語や構文を検索・抽出できるコーパスをインタフェースとともに構築することを目標としている．

データとしては，新聞記事をはじめとして，構文のしっかりした現代の書き言葉を取り上げる．この点，東北圏のブロック紙『河北新報』を発行する河北新報社の協力を得られることになっている．

アノテーションのスキーマとしては，汎用性を重んじて中立的なものを採用し，特定の形式言語理論にはコミットしない．アノテーションの方針決定と実際の作業に当たっては，検索に利用可能か否か，また一貫性および正確性を基準として行う．

3　なぜ句構造か？

句構造情報をタグ付けしたコーパスを開発する理由は，統語構造の曖昧性を克服して，正確な意味情報を抽出するためである．例えば，次の (1a, b) はともに，名詞「写真」を関係節が修飾する構文である．

(1) a.　昨日とった**写真**
　　b.　子供が泳いでいる**写真**

しかし「写真」は (1a) では関係節の述語「とった」の直接目的語としての役割を果たしている (寺村 1975 の言う「内の関係」) のに対し，(1b) ではそのような格役割を関係節の中で果さない．関係節は「写真」の内容を示している (「外の関係」)．本コーパスでは，このような違いを図 1 に示すアノテーションの違いとして表す．図 1a では関係節は IP-REL のタグを与えられ，その内部にあらわれるトレース *T* が直接目的語 (NP-OB1) として扱われている．これに対して，図 1b の関係節は，埋め込みを表す IP-EMB とタグ付けされ，トレースは表れない．それぞれ IP-REL および IP-EMB を手掛かりとして，図の最下部の論理意味表示が自動生成される．「内の関係」の図 1a では 2 つの述語「写真」と「とる」は連言結合子 ∧ で結び付けられているのに対し，「外の関係」の図 1b では「泳いでいる」の意味に相当する意味が「写真」の意味の中に埋め込まれている．

文節と形態論情報にもとづく既存のコーパスでは，上記の (1a,b) は同様のアノテーションを与えられ，区別されていない．文節コーパスの一部には各述語の格フレーム情報をアノテートしたものがある (小原 2013) ことから，これを利用すれば「内の関係/外の関係」の区別が可能であると主張する人がいるかも知れない．関係節を構成する述語の必須格と主名詞とがマッチす

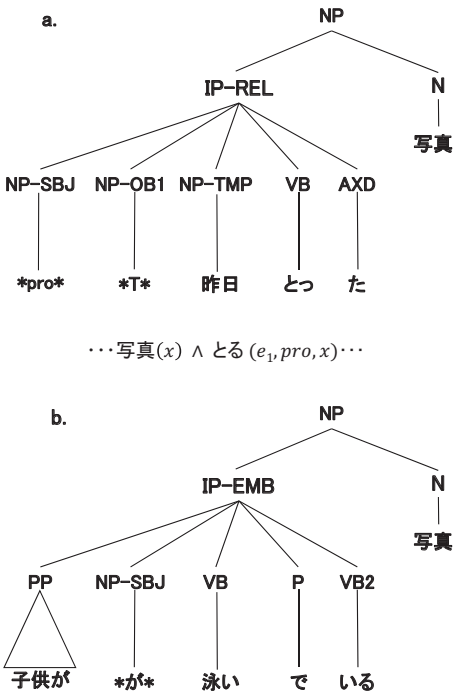

図 1: 文 (1a,b) の統語解析

るなら「内の関係」，そうでないなら「外の関係」だというのである．しかし，この説は成り立たない．第一に，日本語には，

(2) リンゴをむいた**ナイフ**

のように，道具格を含む任意格となるトレースを持つ関係節の例がある．各述語について，任意格をも含めた格フレームのデータベースを作ることは事実上不可能である．第二に，関係節は，非有界依存構文 (unbounded dependency) の一種であり，トレースと主名詞とがどれだけの語句によって隔てられるかはあらかじめ予測がつかない．適切な統語または意味解析を行うのでなければ，両者の依存関係はつき止められない．第三に，そもそも述語の格フレームは固定したものでなく時と場合により変容するものであることが最近の研究でも明らかにされており，データベースによるアプローチには限界がある (Sasano et al. 2009).

4 なぜ句構造＋論理意味表示か？

上記のように，関係節中の述語と主名詞とは非有界依存の関係にあり，下の (3a-c) における「買う」と「メロン」のように，その間がどれだけの長さの語句によって隔てられるか分らない.

(3) a. 私が 買った **メロン**
b. 私が 買って，冷蔵庫に入れてある**メロン**
c. 私が 買って，冷蔵庫に入れてあると思った **メロン**

これまでに公開されたすべての日本語コーパスにおいて (他の言語についてもほぼ同様であるが)，このような非有界依存関係はアノテートされていない．そのため，格支配関係にある動詞と名詞句とを抽出したい場合，せいぜい全体の一部しか情報が得られず，非有界依存関係にあるものはカバーできない.

非有界依存関係へのアプローチとしては，従来は形式統語理論を使って複雑な構文の解析を行う手法が中心であったが，これは少数のサンプル文を対象とする実験を除いて成功していない．これに対して，バトラーの提唱するスコープ制御理論 (Scope Control Theory, 以下 SCT; Butler 2010 参照) では，自然言語の文構造は意味の基本的な構成要素であるスコープ (変項の値) を導入したり，操作あるいは評価をすることによって形作られる依存関係を直接反映したものであると考える．統語構造としては単純な表層統語構造のみを仮定し，スコープを制御することにより，文の意味解析 (評価) を行う．これによると，上記の非有界依存関係に関する情報は，統語解析のレベルでは与えられなくともよい．実際，図 1a では，トレース *T* が主名詞「写真」と同一であることを示すインデクスは付加されていない. この統語情報を SCT をインプリメントした意味解析システムに入力することによって論理意味表示が得られ，この中で両者の情報がリンクされている (. . . 写真 $(x) \wedge$ とる (e_1, pro, x) . . .) ので，結果として依存関係がつき止められるのである.

さらに，従属節中で主語が明示されず，それを埋め込む上位の節の主題や主語等と一致する文が日本語には多く見られる．SCT によればこれらの例は，ゼロ代名詞としてではなく埋め込む節の主題，主語等によるコントロールとして扱うことが出来るが，これはより日本語の実情に合った処理であると考えられる．下の (4) で,

(4) **花子は** 果汁を 凍らせ て，デザートを 作った.

文の主題 (同時に主語でもある)「花子は」は，従属節の述語「凍らせ」および主節の述語「作った」によって共有されている．これらの「花子–凍らせる」や「花子–作った」間におけるような依存関係も，SCT にお

けるアノテーションを通じて初めて把握することが出来るのである．

5　構築方法

図 2 に本コーパス開発の過程を示す．テクストは①で，MeCab,UNIDIC および Comainu を使用して形態素解析に掛けられる．名詞句は長単位，述語句は短単位にもとづく解析が原則である．その結果は②で，Berkeley Parser および Bitpar Probabilistic Context-Free Grammar にもとづく統語解析を受ける．その解析結果は多くの誤りを含むので，人手による修正が必要である．また，修正結果は統語解析プログラムへとフィードバックされる．修正を経た統語解析結果は (図 3 に例文 (4) の解析例を示す) は，Awk/Perl で作成した構造変換プログラム (③) により，SCT 中間表示に変換される．これが Standard ML で作成した意味解析システム (SCT のインプリメンテーション，④) に入力され，論理意味表示を出力する．意味表示の例を図 4 に示す．

図 4 に示すような論理意味表示からは，さらに有用な様々な意味情報を抽出，加工することが出来る．その利用法としては今後の研究の必要があるが，ここでは文法に関連する事柄から 1 点のみ取り上げる．第 4 節に述べたように，コントロール構文や非有界依存構文における依存関係，ここでは特に格名詞句と述語の依存関係は，SCT によって意味的関係として捉えることが可能になった．これを論理意味表示として示すことも出来るが，より直接的に，コンピュータ・システムのスタック・トレースのデータを表示して管理するための手段である Flame Graph を利用して視覚化することも可能である．図 5 に文 (4) の語句間の依存関係を視覚化した図を示す.「せ_た」と「作っ_た」とがともに主語 (arg0) として「花子」を共有している．この方法により，テクストデータ中に出現するすべての用言 (動詞，形容詞，助動詞) の格フレームのリストを生成することが出来る．このような基本的でしかも有用な情報を自動的に得られるのは初めてのことである．

6　アノテーションの方式

アノテーションの方式としては，ペン通時コーパス (Penn Historical Corpus; Santorini 2010) のものを採用する．これは世界の多様な言語のコーパス開発に採用されており，世界の研究者にとって利用しやすく，

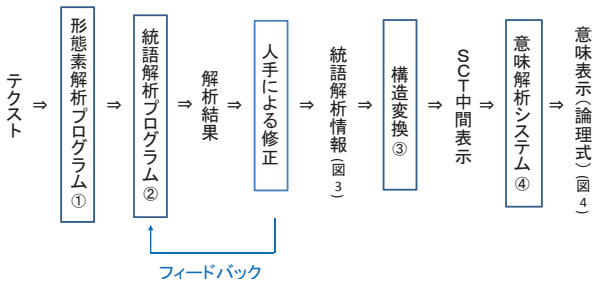

図 2: コーパス開発の過程

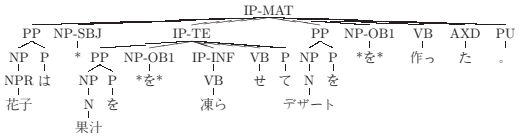

図 3: 統語解析情報の例

$\exists x_1 x_2 e_3 e_4 e_5$(
果汁（x_1）∧
デザート（x_2）∧
て（せ（e_4，花子，x_1，凍ら（e_3，x_1）），作っ_た（e_5，花子，x_2）））

図 4: 論理意味表示の例

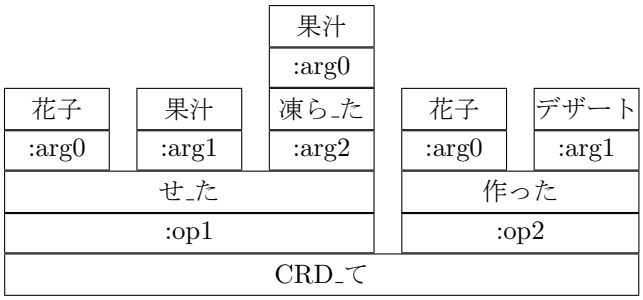

図 5: 格フレームの例

また他の言語のコーパスとの比較や対照が容易だという利点がある．さらに，他の言語における多様な文法現象の取扱いを本コーパス構築に当たって参考にできる．

語彙ラベルとしては，記号を除いて 23 種類，句のラベルとしては 17 種類がある．句ラベルのうち，CP, IP, NP は，さらにハイフンに続けて機能タグを追加できる．これは，統語構造の曖昧性を克服し，正確な意味情報を抽出するために利用される．図 1 の a と b を IP-REL および IP-EMB で区別するのはその一例である．また，必須格名詞句には NP-SBJ, NP-OB1, NP-OB2 のノードが伴うが，これは日本語の主語・直接目的語・間接目的語の表示が曖昧で文法的格を明示

する必要があるためである．

統語解析としては，X-bar 理論を採用して，すべての種類の句が同一の，比較的フラットな構造を持つ．ヘッドはつねに句の右端にあらわれ，句のレベルとの間に中間的なノードは存在しない．修飾語句 (modifiers) や補語句 (complement) は，ヘッドと同一レベルの姉妹となる．これにより木構造の検索や変換が簡単に行われる．また，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．これにより，出現順位の早いものほど広いスコープを持つというディフォールト設定を基本とする，柔軟なスコープ包含関係の指定を可能にする．

4 節で述べたように，関係節中のトレースなど，標準的な統語解析においてはインデクスを用いて行われる指示対象の同定は，SCT による意味解析のレベルで行われる．このため，統語解析におけるインデクスの付加は不要となり，非有界依存構文を含む複雑な構文の依存関係を提供できるにもかかわらず，コーパス構築は現実的な作業量の範囲内で行えることになる．

7　特色と意義

本コーパスは，十分な量の日本語データについて，正確な統語・意味解析情報，特に複雑な構文も含めた語句の間の依存関係に関する情報を提供する最初のコーパスである．依存関係にもとづいて，必要な構造体をピンポイントで検索できる．検索用インタフェースとして，正規表現を利用するもの以外に，メニュー方式やグラフィカルユーザインタフェースを用いた，初心者用のものも提供する予定である．

上記のように，ペン・ツリーバンクに共通するアノテーション方式を採用し，また英語によるマニュアルとインタフェースも提供するので，日本語に関して未熟な海外の研究者にも日本語文法に関する情報を提供できる．

本コーパスは，日本語文法および日本語に関わる関連分野，たとえば語用論、テクスト分析、社会言語学，言語心理学，神経言語学の研究者に対して，言語データにもとづく客観性と理論的正確性を兼ね備えた研究を行うための手段を提供する．自然言語処理研究における利用は言うまでもない．さらに，異なる言語のコーパス間の比較にもとづくコーパス言語類型論など，これまでに無い新分野の開拓へとつながることが期待される．

8　予定される成果と将来の展望

将来的には数万文規模の統語・意味解析情報付コーパスを完成させ，検索用インタフェースとともに公開する予定である．アノテーションの蓄積がパーザーの進化へとフィードバックしていくことも期待され，さらにその後，より精緻なアノテーションを施した，英語ツリーバンクの大きさに匹敵する大規模コーパスの構築と公開を視野に入れている．

謝辞

この研究は，大学共同利用機関法人人間文化研究機構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」の支援を得て行われました．

参考文献

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

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 Academic Publishers. 2003.

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

小原京子「日本語フレームネット: 文意理解のためのコーパスアノテーション」『言語処理学会第 19 回発表論文集』, 2013.

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.

Sasano, R., et al. The effect of corpus size on case frame acquisition for discourse analysis. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACLS.* 2009.

寺村秀夫「連体修飾のシンタクスと意味その 1」『日本語・日本文化』4, 大阪外国語大学留学生別科, 1975.