「統語・意味解析コーパスの開発と言語研究プロジェクト：現状と今後の展望」

プラシャント・パルデシ（国立国語研究所）、窪田悠介（筑波大学）

1.概要

文の統語解析情報（句構造）を付加したツリーバンクの開発は、海外のコーパス開発では重要な研究の一つと位置付けられている。2016 年 4 月から国立国語研究所で文の統語解析情報（句構造）をアノテートした NINJAL Parsed Corpus of Modern Japanese (NPCMJ) の開発が進められ、2018 年 3 月現在 2 万文（約 31 万語）規模のコーパスが、①タグ・ブラウザー、②文字列検索、③ツリー検索とテキスト解析、④クエリー作成の 4 種類の検索ツール（インターフェース）とともに無償一般公開されている。本コーパスはペン通時コーパス (Penn Historical Treebank) のアノテーション方式を採用しており、統語情報のアノテーションは表層的､中立的なものであり､特定の形式言語理論にコミットしていない。

本発表の前半では NPCMJ 開発の現状と今後の展望を述べ､このコーパスを手軽にブラウズできる開発中の検索ツール「NPCMJ Explorer」を紹介する。後半ではこのコーパスに基づく言語研究の事例を報告し言語理論研究における統語構造付きコーパスの有用性を示す。

2.NPCMJ コーパスおよびその検索ツールの開発・公開：「NPCMJ Explorer」

一般に、統語情報を付加したコーパスを検索するにはツリー構造を検索することのできるクエリー言語を利用する必要がある。しかし、クエリー言語の習得には時間がかかり、これが統語情報付きコーパスを利用する上での大きな障害となる。そこで、ユーザが検索式を一切入力することなく項目名をクリックするだけで、日本語の主要な文法項目を含んだ NPCMJ の用例文やその構文構造を簡単に確認することができるツール「NPCMJ Explorer」を開発している。検索項目の選定にあたっては、益岡隆志・田窪行則（1992）『基礎日本語文法』（くろしお出版）を利用し、この文法書にある 136 の節項目のうち、73 項目をカバーした｡本発表では 2019 年 3 月公開予定のこのツールの現時点のバージョンを紹介する。

3.NPCMJ に基づく言語研究

統語構造付きコーパスの言語研究への応用として､Kubota & Kubota (2018) を紹介する。日本語のテ形節、連用節に関しては、これらの構文を等位接続構造と考えた場合に等位接続構造制約の反例となるような、(1) のような例が存在する (Kubota & Lee 2015) 。

（1）これが太郎が [紀伊国屋に行って/行き] [__ 買った] 本だ。

このような例文が実データに現れるかを効率よく調べられるかは、理論的な立場に関わらず統語論、意味論研究者にとって重要な問題である。Kubota & Kubota (2018) は NPCMJ を用いて（1）に類する実例が容易に検索可能であることを示した。また、コーパスからデータを網羅的に取ることで見えてくる傾向性が、文法研究に有用であるという結果も得られた。