W-5

統語・意味解析情報付き日本語コーパスの構築に向けて

企画者: プラシャント・パルデシ (国立国語研究所)
司会者: 吉本 啓 (東北大学)
コメンテーター: 福島一彦 (関西外国語大学)

prashant@ninjal.ac.jp， kei@compling.jp， kaz.zisin@gmail.com

1　ワークショップの趣旨

従来公開されてきた日本語コーパスは、文の文節への分析を基礎として、形態素解析情報および文節間の係り受け関係をタグ付けしたものが中心であった。しかし、文の意味を直接反映する構造は句構造であり、言語学者が必要とする深い情報を文節とその係り受け関係から自動的に得ることには限界がある。そのことから、文の統語解析情報 (句構造) をアノテートした日本語コーパス (ツリーバンク) の開発に着手した。このコーパスは、各文に対して意味解析情報 (論理意味表示) が付加されているという特徴を持つ。このワークショップでは、コーパス開発の動機、特色、アノテーション方式、および意義について解説し、具体例のデモンストレーションを行ってコーパスについて理解を深めるとともに、参加者からのフィードバックを得て今後に役立てることを目的としてワークショップを行う。

2　各発表の題目、発表者氏名および要旨

(1) イントロダクション

発表者: プラシャント・パルデシ

日本語統語・意味解析情報付きコーパス開発の動機およびプロジェクト概要について説明する。本コーパスのアノテーション方式としては、ペン通時コーパス (Penn Historical Treebank; Santorini 2010) のものを採用している。この方式は世界の多様な言語のコーパスに利用されていることから、言語間の比較や対照が容易であり、コーパスを利用した対照研究や類型論研究を可能にする。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。6 年間のプロジェクト期間中に 5〜6 万文からなる日本語コーパスを構築する予定である。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。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。

(2) アノテーション方式とコーパスの特色

発表者: 吉本啓

ペン通時コーパスのアノテーション体系に従いながら、日本語の実情に適合し、論理意味表示を出力するという目的を果たすために必要なアノテーション方式を編み出した。句に付けられるラベルには、必要に応じ機能表示が付加される。さらに、必須格名詞句が省略された文に対しても、ゼロ代名詞のタグ付けにより正確な統語・意味解析情報を提供する。また、各文に対しタグ付けされる意味解析情報にもとづいて、文を構成する語や句の間の統語・意味関係 (依存関係) がすべて明示されるが、このことは様々な文法情報の自動的な獲得を可能にする。このように、本コーパスは、十分な日本語データについ

て複雑な構文も含めて正確な統語・意味解析情報を提供する最初のコーパスであり、簡便なインタフェースの開発により、言語情報処理技術に通じていない研究者にも構造体をピンポイントで検索することを可能にする。さらに、コーパス拡張のための XML エンコーディングおよび開発支援ツールについても説明する。

(3) デモンストレーション

発表者: アラステア・バトラー、窪田愛、窪田悠介

本プロジェクトにおいて開発を行っている、コーパスの利用を容易にするためのユーザー・フレンドリなインタフェースについて解説する。まず、構築されたコーパスの表示において、テクスト中のすべての文について、形態論・統語論情報がリンクされていて簡単に表示でき、また前後の文脈についても簡単に知ることが出来ることを示す。次に、入力された語句に相当するデータを形態・統語論情報や頻度情報とともに KWIC 形式で出力するインタフェースを開発しており、ある語句の形態・統語論的な用法について知るのに利用することが出来る。最後に、あらかじめ与えられたサーチ・パターンを利用して、関連する構造体 (constructions) や統語情報を検索するパターン・ブラウザがある。一度使用した木構造にユーザーが手を加えることによって、別の関連する統語パターンの検索が可能になり、強力なツールを提供する。

(4) まとめと将来の展望

発表者: プラシャント・パルデシ

各々のプレゼンテーションについて総括を行うとともに、本コーパスプロジェクトが日本語研究全体において持つ意義を述べる。さらに、将来にわたるコーパス情報の拡張についても展望を述べる。

参考文献

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.