基幹型共同研究プロジェクト

「コーパス日本語学の創成」：コーパスを用いた日本語研究の推進方策と実績
「コーパスアノテーションの基礎研究」：アノテーションの重ね合わせ技術を中心に

前川 喜久雄

《研究の概要》

「コーパス日本語学の創成」（以下「創成」と略）はコーパスを利用した日本語研究を定着させるためのメタプロジェクトである。共同研究者の枠にとらわれず、広く一般に研究発表の場を提供するとともに、コーパスの設計・構築に関わる基礎知識の普及のために一連の書籍を刊行しつつある（共同研究者は 43 名）。「コーパスアノテーションの基礎研究」（以下「アノテーション」）では『現代日本語書き言葉均衡コーパス』などの既存コーパスの研究上の価値を向上させるために、各種アノテーション技術の研究を進めるとともに、アノテーションデータの利用環境についても研究開発を進めている（共同研究者は 20 名）。

《主要な成果物》

「創成」

1. 日本音声学会誌『音声研究』特集「大規模コーパスを利用したデータ駆動型音声研究」（募集中）
2. 『コーパス日本語学ワークショップ予稿集』（第 1 回～第 5 回）（2011～2014）
3. 『コーパス入門』(講座日本語コーパス第 1 巻) 朝倉書店, 2013. （全 8 巻予定）
4. *A Frequency Dictionary of Japanese: Core vocabulary for learners.* London: Routledge, 2012.
5. 『現代日本語書き言葉均衡コーパス』(DVD 版) 国立国語研究所コーパス開発センター, 2011.

「アノテーション」

1. 言語処理学会誌『自然言語処理』特集号「コーパスアノテーション―新しい可能性と共有化にむけての試み―」（2014 年 4 月刊行予定）
2. 『現代日本語書き言葉均衡コーパス』構築関連マニュアル類の整備と公開（2013）
3. 言語処理学会テーマセッション「コーパスアノテーションの可能性と共有化」（2012）

《特色ある活動》

「創成」では、コーパスを用いた日本語研究に興味をもつ研究者の交流の場として、毎年 9 月と 3 月に 2 日ずつ「コーパス日本語学ワークショップ」(JCLWS)を開催しており、2014 年 3 月が第 5 回となる。日本語学にとどまらず、情報処理系、認知科学系の発表も多い。変わったところでは図書館学、法律学などを専門とする研究者の発表もある。

毎回 50 件前後の発表があり、その約半分が国語研とは直接関係のない研究者（共同研究者以外の）による発表であり、大学院生の発表も全体の四分の一前後を占めている。予稿集は毎回 400 頁前後になるが、すべて国語研のホームページからダウンロードできる(URL は後掲)。毎回 200 名前後の参加者があるので、実際上、コーパス日本語学に関する学会の機能を提供できていると考えている。

「アノテーション」では言語処理学会の査読付論文誌『自然言語処理』の特集号「コーパスアノテーション―新しい可能性と共有化にむけての試み―」の編集を進めている。

これは 2012 年度に実施した学会のテーマセッション「コーパスアノテーションの可能性と共有化」が発展した特集であり、従来、論文になりにくいといわれてきたデータ構築関係の仕事を研究として正面からとりあげようとするところに、特色がある。

国語研関係者の提案になる特集であるが、招待論文のような形で、共同研究のメンバーを特別扱いすることはしないので、メンバーも通常の会員同様、査読を通過しなければ論文は掲載されない。

この特集には言語資源研究の様々なテーマについて 14 篇の投稿があり、現在審査の最終段階にある。本年 4 月の刊行を予定している。

《何が分かったか，何が出来たか》

今回のプレゼンでは「アノテーション」の成果に焦点をあてる。本プロジェクトが主なアノテーション対象としている『現代日本語書き言葉均衡コーパス』(BCCWJ)には、文書構造タグと短単位と長単位による形態論情報（形態素解析結果）とが提供されているが、一層多くのアノテーションを提供することで、研究の可能性が飛躍的に広がると考えられる。

本プロジェクトでは、以下に挙げる情報について、実際の作業を通じてアノテーションの仕様を決定してマニュアルを作成し、最終的にはマニュアルとデータ（BCCWJ の一部に対するアノテーション）を公開することを目標に共同研究を進めている。

文の構造に関わる①文節係り受け構造のアノテーション。②述語項構造（含、共参照関係）、③日本語フレームネット、④動詞項構造シソーラスのように述語を中心とした文の内部構造に関わるアノテーション。⑤拡張固有表現、⑥時間情報表現、⑦助動詞「れる・られる」の意味、⑧述語境界位置のように文中のセグメントに関するアノテーション。⑧拡張モダリティ、⑨否定の焦点のように、文構造とセグメントの中間レベルのアノテーション。さらに、書き言葉を黙読する際の読み時間の情報や話し言葉の持続時間長やイントネーションのような連続量も研究対象の一部に含めている。

ところで、このように多様なアノテーションが提供されるようになると、各アノテーションを本来の研究目的に利用するだけでなく、複数のアノテーションを有機的に関連づけた研究の可能性が拓けてくる。例えば、特殊な句末イントネーションの機能を知りたいときに、形態論情報を利用した節境界の分類とそのイントネーションの生起確率を関連づけて分析し、生起確率の高まる節境界の文法的特徴からイントネーションの機能を推測するというような研究である。

しかしアノテーションの重ね合わせは簡単でない。これまでに開発されたなかで最も複雑なアノテーション情報をもったコーパスは『日本語話し言葉コーパス』（CSJ）のコア部分であろうが（形態論情報、節境界にくわえて、分節音、韻律、談話構造などのアノテーションが施されている）、CSJ ではすべてのアノテーション情報を一つの XML ファイルにまとめて格納したために、データの構造が過度に複雑化してしまい、CSJ コアの XML ファイルを分析に利用できる研究者の数が―特に文科系において―きわめて限られたものになってしまった。

この反省にたって、我々は現在、種々のアノテーションのうち、研究者が必要とする部分だけを選択的に組み合わせて利用可能なソフトウェア環境の開発を試みている。具体的には二つの可能性を試行している。ひとつは従来から奈良先端科学技術大で松本裕治氏と浅原正幸氏（現国語研）を中心に開発が進められてきたコーパス検索・アノテーションツール ChaKi.NET の利用、もうひとつは近年、狩野芳伸氏（国立情報学研究所）が開発している統合自然言語処理システム Kachako の利用である。

ChaKi.NET は書き言葉コーパスを利用するために用いられるツールで、内部で形態素解析・係り受け解析器を呼びだすことでテキストを解析し、形態論情報・係り受け構造をもとに調査対象である文を絞り込み検索することができる。入出力に拡張 CaboCha 形式を用い、文の一部を切り出すセグメント、セグメント間の有向関係を表すリンク、セグメントの同値類を表す同値類を表すグループなどが表現でき、これらを可視化することができる。今後複数のアノテーションの重ね合わせを行う際に発生するアノテーションの名前衝突を回避するための機構や、各種アノテーションを組み合わせて問い合わせるユーザインターフェイスの改良に努める。

一方、Kachako は大量の非構造化情報を分析し、その中から有益な情報を発見、構造化してエンドユーザーに渡すアプリケーションを開発するためのソフトウェアフレームワーク UIMA (Unstructured Information Management Architecture) に準拠した統合実行基盤およびソフトウェアコンポーネント群である。形態素解析や係り受け解析器を呼び出すだけでなく、多様なアノテーションを読み込む機構を備えている。今後、各種アノテーションを読み込み、検索する機構を開発する予定である。

先に述べたように、共同研究の終了までには、各種アノテーションデータをマニュアルとともに公開し、ChaKi.NET ないし Kachako で自由に利用してもらえる環境を実現したいと考えている。

参考 URL

JCLWS: http://www.ninjal.ac.jp/research/project/a/sousei/

ChaKi.NET: http://sourceforge.jp/projects/chaki/

Kachako: http://kachako.org/kano/