コーパスアノテーションと心理言語学

浅原 正幸（国立国語研究所コーパス開発センター）†
小野 創（近畿大学理工学部）
狩野 芳伸（科学技術振興機構さきがけ）

Corpus Annotation and Psycholinguistics

Masayuki Asahara (Center for Corpus Development, NINJAL)
Hajime Ono (School of Science and Engineering, Kinki University)
Yoshinobu Kano (PRESTO, Japan Science and Technology Agency)

シンポジウム「コーパスアノテーションと心理言語学」開催趣意

言語のカタチをとらえるために数多くのコーパスが構築され共有されてきた。日本語においては『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）が完成し、昨年末より公開が始まった。今後コーパスを用いて言語運用を定量的に評価する日本語研究が盛んになることが期待される。また、言語処理研究者を中心に、コーパス上に、データの出自などのメタデータ以外に、形態論情報や統語意味論情報などのアノテーションが付与されてきた。これまでは、新聞記事など共有化できるコーパス上にアノテーションを重ね合わせてきたが、今後、他の多様なジャンルを含む BCCWJ 上に異なるレベルのアノテーションを重ね合わせ、共有されていくことだろう。

コーパスのアノテーションは、コーパスコンコーダンサなどを介して検索や統計量取得の際に利用され、言語学研究者による調査のための手がかりとして用いられる。また、言語処理研究者による言語解析器の開発のための教師ありデータとして用いられ、さらに解析手法の性能評価のためのベンチマークデータとしても用いられる。いずれの分野の研究者も、コーパスのアノテーションに対し、誤りやゆれが少ないことを求める。アノテーションに携わる者は基準を検討したり、ツールを整備したりすることにより、アノテーション誤りを少なくし、複数人のアノテータ間のゆれだけでなく 1 人のアノテータによる時間経過によるゆれを少なくする努力を続けてきた。しかしながら、この問題について決定的な解決方法は未だ提案されていない。

ここで、誤りやゆれの原因について、アノテーション過程の観点から考えてみたい。アノテーションのないコーパス（生コーパス）はテキストの原著者による言語の生成過程の産物であるが、アノテーションそのものは生コーパスのテキストを読むアノテータの受容過程の産物である。テキストの生産者（書き手）と受容者（読み手）との間に認識の齟齬があり、このことがアノテーションの誤りやゆれをもたらす。生コーパスそのものを調査することは、言語の生成過程をとらえることにほかならない。これに対して、アノテーションを調査することは言語の受容過程をとらえることであり、書き手と読み手の間にある個人の経験の差という要素に左右され誤りやゆれが必然的に入り込む。このような観点を考慮すべきである。

心理言語学の分野では、人間の心理的な観点から、人間が言語を獲得する過程、言語の生成過程と受容過程をそれぞれ研究対象としてきた。研究の方法論として、文生成課題、自己ペース読解課題、視線追跡読解課題をはじめとして、近年では脳波計を用いた事象関連電位の調査など被験者実験を中心としており、社会的に共有される言語規範だけでなく、共有されない個人の経験に基づく言語規範についても明らかにしてきた。これに対して、従前のコーパス言語学では、新聞など統制されたテキストを調査対象とし、社会的に共有されている言語規範を明らかにしてきた。今後、多様な書き手を含む BCCWJ を用いて、

† masayu-a@ninjal.ac.jp

どのような書き手が文法規範から外れた表現を生成するかを調査することにより、言語の生成過程における個人の経験の差という要素を明らかにすることができるようになろう。

ここで、コーパスアノテーションを言語の受容過程と考え、心理言語学的な観点から見つめなおすことを考えたい。コーパスアノテーションにおいて、基準やツールを用いて、社会的に共有されている言語規範については統制して誤りを減らす努力はすべきである。しかしながら、そうではない言語現象に対しては、誤りやゆれを認めたうえで、これを受容過程における個人の認識の差異として分析対象とし、定量的評価を行いたいと考える。

では、具体的にどういったことができるだろうか。ここで 2 つの方法論について考えたい。1 つは、コーパスアノテーション過程そのものを心理言語学における被験者実験ととらえ、アノテーションの誤りやゆれそのものを評価することである。基準やツールを用いても統制できない誤りやゆれについて、アノテータの作業過程で何が起こっているのかを深く検討することができるであろう。もう 1 つは、コーパスに対し心理言語実験で用いられている手法を用いて可読性などを定量評価した情報を付与し、その情報をコーパスアノテーション作業に生かすことである。適切にサンプリングされたテキストデータに対して、複数人による心理言語実験を行い、リーダビリティの低いテキストを定量的に評価して、その情報をもとにアノテーション誤りを検出することができるであろう。

言語解析器の開発の立場からは、定量的に評価された誤りやゆれの情報をどう扱えばよいだろうか。例えば、識別モデルに基づく機械学習手法を用いて、社会的に共有されている言語規範として集積されたアノテーションを正例とし、アノテータによって生成された明らかなアノテーション誤りを負例として推定することが考えられる。また、生成モデルに基づく機械学習手法を用いて、アノテータ間の受容過程のゆれを考慮して適切な分布を推定することが考えられる。

本シンポジウムでは、浅原、小野、狩野の 3 人の若手研究者より上に述べた観点から問題提起を行う。浅原が上に述べた問題意識について説明し、小野が実験言語学的な研究の方法論について概説し、狩野がコーパス言語学と心理言語学を結びつける方法について話題提供する。

登壇者プロフィール

浅原正幸（国立国語研究所コーパス開発センター特任准教授）

奈良先端科学技術大学院大学情報科学研究科博士後期課程修了。博士（工学）。学術振興会特別研究員（DC1）、奈良先端科学技術大学院大学情報科学研究科助手、助教を経て、2012 年 1 月より現職。形態素解析器、構文解析器の開発と言語処理向けのコーパス整備に従事。現在はコーパスに対する統語意味論構造のアノテーションと言語研究向けの超大規模コーパスの設計・開発に従事。

小野創（近畿大学理工学部専任講師）

米国メリーランド大学人文学研究科博士課程修了。Ph.D（Linguistics）。広島大学特別研究員、関西外国語大学特任講師を経て、2009 年 4 月より現職。統語論を中心にした理論言語学の研究､特に日本語と英語の感嘆文の wh 依存関係に関する統語･意味論研究に従事。近年は、日本語かき混ぜ文の文処理（行動実験や脳波計測実験）、疑問詞などの wh 依存関係の処理、受身文の産出、存在文の産出と視線の関係などの研究に従事。

狩野芳伸（科学技術振興機構さきがけ研究者・国立情報学研究所外来研究員）

東京大学情報理工学系研究科博士課程単位取得退学。博士（情報理工学）。東京大学情報理工学系研究科特任研究員等を経て、2011 年 10 月より現職。言語資源（ツール・コーパス）の互換性と相互運用性の研究、相互比較や視覚化を行う統合環境の設計・開発に従事。現在はさらに大規模処理や機械学習を統合した自動化システムと、心理学的妥当性を考慮した言語処理モデルとその応用の研究に従事。