１．研究領域名：代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備
２．研 究 期 間：平成 18 年度～平成 22 年度
３．領域代表者：前川　喜久雄（独立行政法人国立国語研究所・研究開発部門言語資源グループ・グループ長）

４．領域代表者からの報告

（１）研究領域の目的及び意義

本研究領域の目的は、従来、諸外国に比べて立ち遅れていた日本語コーパスの整備を進めると同時に、コーパスを利用した日本語研究の方法を開拓することによって、今後数十年にわたる日本語研究の基盤整備を行うことにある。本研究領域は、同時期に実施される国立国語研究所の KOTONOHA 計画と連携して、合計で 1 億語規模の『現代日本語書き言葉均衡コーパス』を構築し公開するが、そのうち書籍データ約 5000 万語分が本研究領域の分担である。

研究項目 A01 では上述の書籍コーパスを構築し、そのために必要とされる電子化辞書や各種アノテーションツールを開発する。このコーパスは著作権処理を施して、研究期間終了後に一般公開する。研究項目 B01 では構築したコーパスを利用した言語研究を基礎と応用の両面において実施し、コーパスに依拠した言語研究の方法を開拓するとともに、研究項目 A01 で開発したコーパスの有用性を評価する。言語学（日本語学）的な基礎研究に加えて、日本語教育学、言語政策（国語教育を含む）、辞書編集、自然言語処理（特に意味処理）の各領域に計画研究を設置して研究を推進し、研究の公募も実施する。

本研究領域の意義として、日本語のいわゆるコーパス言語学的な研究が活性化されるのは当然であるが、さらに、言語教育や言語政策など、従来ともすれば実証的なデータを欠いた議論が行われがちであった領域でも、実証的な研究活動が行われるようになることが期待される。

（２）研究の進展状況及び成果の概要

研究期間前半では研究項目 A01 を中心にコーパス構築を急ぎ、後半での研究項目 B01 の研究活動を円滑化させることが基本戦略なので、以下では主にコーパスの整備状況について報告する。サンプリングと入力は当初予定を 1 割程度上回るペースで進捗しており、2008 年 7 月時点で、書籍（3300 万語）、白書（500 万語）、インターネット掲示板（500 万語）、国会会議録（500 万語）など、5000 万語以上のデータを利用できる。このうち著作権処理が完了した 2800 万語分は、領域外の研究者にも公開しており、多数の利用者がある。

データの形態素解析用辞書である UniDic には、斉一性の高い言語単位として我々が提案する短単位に則って約 12 万語が登録されている。既存の形態素解析ソフトを UniDic で利用すると、新聞や白書の場合、精度 98%以上での自動解析が可能である。UniDic も一般公開しており、1000 件以上の利用がある。

研究項目 B01 の代表として、言語政策関係の活動をとりあげて紹介すると、書籍コーパスを用いて国語教育用基本語彙の選定の目安を得たこと、文化審議会国語分科会による常用漢字表見直し作業で生じた問題を解決するためにコーパスから抽出したデータを提供したこと、さらに独自に検定教科書のコーパス（500 万語）を整備したことなどの成果があがっている。B01 の他研究項目も活発に活動しており、領域全体で報告書 18 冊（合計 2915 ページ）、論文 91 件（うち査読付 27 件）、学会発表 222 件（うち国際学会 56 件）、受賞 4 件の実績があがっている。

５．審査部会における所見

Ａ　（現行のまま推進すればよい）

本研究領域が目的に掲げた大規模日本語書き言葉コーパスの構築は、今後の日本語研究の基盤のひとつであることは言をまたず、これが特定領域研究において推進されていることのもつ意義は高い。コーパス構築は当初の計画よりもその進度は速く、さらに、構築されているコーパスは質的な面においても他国で構築されているコーパスを凌駕するものと判断する。

各計画研究の研究進展状況も順調であり、コーパス構築に必要な解析ツールを領域内の計画研究で開発する等、それぞれの特長を活かした研究成果が領域内で有機的に連結している点において、領域の目的、成果への見通しが各研究課題において共有されているものと判断する。また、これらの研究について、本邦及び海外における成果発表が積極的になされており、学術的意義は高い。

一方で、今後の課題として求められていることは、専門研究者でない方々、日本語話者以外の方々にも利用が容易になるような配慮を加え、構築したコーパスの継続的な活用のための見通しを明らかにしていくことである。これらの点に留意しながら、本研究領域の研究を今後も推進していくことを望む。