様式 C-7-2

自己評価報告書

平成２１年５月２７日現在

研究種目： 特定領域研究
研究期間： 2006～2010
課題番号： １８０６１００９
研究課題名（和文） 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
研究課題名（英文） Compilation of a Balanced Corpus of Written Japanese:
Infrastructure for the Coming Japanese Linguistics.

研究代表者
前川 喜久雄（MAEKAWA KIKUO）
独立行政法人国立国語研究所・研究開発部門・グループ長
研究者番号： ２０１７３６９３

研究分野：言語学、日本語学
科研費の分科・細目：言語学、情報学・日本語学、知能情報学
キーワード：(1)均衡コーパス (2)日本語 (3)書き言葉 (4)代表性

１．研究計画の概要

本研究領域にはふたつの大きな目標がある。ひとつは、現代日本語のコーパス言語学的研究の基盤を整備するために、大規模な現代日本語書き言葉の均衡コーパスを構築することである。本領域におけるコーパス構築は、国立国語研究所の近現代日本語コーパス整備事業である KOTONOHA 計画と連携して実施する。両者が協力して１億語を超える規模の『現代日本語書き言葉均衡コーパス』を構築するが、分担関係として、本領域では、書籍に用いられた現代語の書き言葉を対象とする 5000 万語規模の書き言葉コーパスを構築して公開する。

本領域のもうひとつの目標は、構築途上のコーパスを様々な領域で利用することによってコーパス日本語学の可能性を探り、同時に構築中のコーパスの有用性を評価することである。狭義の言語学だけでなく、国語教育、日本語教育、辞書編集、自然言語処理などの幅広い領域で活用と評価をおこなう。

２．研究の進捗状況

研究申請時の計画通り、研究期間前半はコーパスの構築に力を注いだ。現時点でコーパスの書籍部分は 5000 万語の予定に対して 4500 万語以上のデータ入力が終了しており、そのうち４割程度は著作権処理も終了している。コーパス全体としては、１億語の予定に対して 7800 万語の入力が終了し、4000 万語以上の著作権処理が終了している。これらはいずれも当初予定を１割程度上回る進捗状況である。

著作権処理まで終了したデータは、その大部分をインターネット上の検索デモンストレーションサイトで一般公開しており、これまでに 50,000 件以上のアクセスがあった(http://www.kotonoha.gr.jp/demo/)。

また平成 19 年度以来、著作権処理済データの大部分を定期的に本領域外の研究者にモニター公開してきているが、現時点で 500 名弱の利用者がある。これは国内外において日本語研究にコーパスを利用する可能性のある研究者の大部分をカバーする数字であると考えられ、本領域の成果に対する学界の期待の大きさを感じさせる。

研究期間後半に入ってからは活動の重点をコーパスの開発から活用に移しながら活動している。研究期間前半においても活発な成果公表がおこなわれてきたが（5 参照）、今後は一層活発な成果公表を期待できると考えている。

これまでに得られた成果のうち、特筆に値する成果をふたつ挙げておく。第一に文化審議会国語分科会では、偶々本領域と並行して常用漢字の見直し作業を進めているが、そこで基礎データとして利用されていた印刷会社提供の活字利用頻度データの妥当性について疑義が呈された。これを解消するために、本領域で開発した書籍コーパスにおける漢字使用実態データの解析結果提供の要請があり、解析結果を提供した。

第二に、従来、国語教育では教育用漢字は定められていても、教育用語彙についての議論が十分におこなわれていない。この問題の解決を念頭におきつつ、本領域言語政策班では、書籍コーパスの解析に依拠して、国語教育用基礎語彙表を試作した。

３．現在までの達成度

①「当初の計画以上に進展している」と判

断する。コーパスの構築が当初の数値目標を１割程度上回っていることと、領域外研究者を含めてコーパスに対する強い興味を喚起していることが判断の根拠である。

実際、平成 20 年度に実施された領域中間評価ではＡ評価をもらい、年度末には追加予算の配分をうけた。

４．今後の研究の推進方策

当初計画に沿って、研究期間後半においては、コーパスを利用した日本語研究に力を注ぐ。その際、本領域内部だけにとどまらず、領域外の研究者をも含め、コーパスを用いた日本語研究の土台を固めることを心がける。

本領域では従来から各研究班が開催する研究会は原則公開とすることによって、班間および領域外との交流を促してきているが、今後もこの方針を維持する。

また 2 で述べたように本領域では領域外の研究者にも著作権処理済のデータをモニター公開してきているが、平成 21 年度には 5000 万語規模のデータをモニター公開する予定である。このモニター版コーパスの利用者は多数に及ぶので、平成 20 年度には、本領域の公開研究成果発表会の前日にモニター版利用者による研究発表を集めたサテライトセッションを開設したところ、大変好評であった。この試みは今年度以降も実施して、本領域の枠を超えた研究者間の意見交換の場とする予定である。

以上のように、本領域はこれまで順調に進捗してきたが、今年度 10 月 1 日には、本領域の中心に位置する国立国語研究所が独立行政法人としては廃止され、大学共同利用機関法人に移管されることが決まっている。そのため、種々の契約や非常勤職員の雇用などの面においての混乱が予想される。有効な対策は存在しないが、できるだけ混乱をおさえるために、管理部門との連携を密にして対応したいと考えている。

５．　代表的な研究成果

領域代表者の業績だけを記載する。分担者の業績は各班の報告書を参照のこと。

〔雑誌論文〕（計 10 件）

① 前川喜久雄「日本語コーパス開発の現状と展望」,『英語コーパス研究』15, pp.3-16, 英語コーパス学会，2008:06.査読無.
② 前川喜久雄「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発」，『日本語の研究』4(1), pp.82-95, 2008:01. 査読無.
③ 前川喜久雄「コーパス日本語学の可能性―大規模均衡コーパスがもたらすもの―」,『日本語科学』22, pp.13-28, 2007:10. 査読無.
④ 前川喜久雄「コーパスとは何か」，『国文学解釈と鑑賞』74-1pp.6-14, 至文堂, 2008:12. 査読無.
⑤ 前川喜久雄・山崎誠「『現代日本語書き言葉均衡コーパス』」，『国文学解釈と鑑賞』74-1，pp.15-25，至文堂，2008:12. 査読無.

〔学会発表〕（計 21 件）

① Maekawa,K. "KOTONOHA: A Corpus Compilation Initiative at the National Institute for Japanese Language", *The 22nd International Conference on the Computer Processing of Oriental Languages(ICCPOL2009)*, 2009:03
② Maekawa, K. "Compilation of the Balanced Corpus of Contemporary Written Japanese in the KOTONOHA Initiative", *Proc. Second International Symposium on Universal Communication*, pp.169-172,2008:12
③ Maekawa, K. "Balanced Corpus of Contemporary Written Japanese." *Proc. The 6th Workshop on Asian Language Resources*, pp.101-102, 2008:01
④ 前川喜久雄「大規模言語資源の開発とその問題点（特に著作権処理について）」WebDB Forum 2008 特別セッション「企業の巨大データ徹底解剖－新たな研究の可能性と産学連携－，情報処理学会，2008:12
⑤ 前川喜久雄「国立国語研究所のコーパス整備計画 KOTONOHA―その現状と問題点―」電子情報通信学会技術研究報告思考と言語 TL2008-1, pp.82-95, 電子情報通信学会，2008:05

〔図書〕（計０件）

〔産業財産権〕

○出願状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ
http://www.tokuteicorpus.jp/