様式 C-19

科学研究費補助金研究成果報告書

平成２４年３月２２日現在.

機関番号：６２６１８
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）
大学共同利用機関法人人間文化研究機構国立国語研究所・言語資源研究系・教授
研究者番号：２０１７３６９３

研究成果の概要（和文）：当初の予定どおりに、5000 万語規模の現代日本語書籍均衡コーパスを構築して 2011 年に公開した。同時に構築途上のコーパスを利用しながら、コーパス日本語学の確立にむけた研究を多方面で推進し、若手研究所の育成にも努めた。現在、約 200 名規模の研究コミュニティーが成立しており、本領域終了後も定期的にワークショップを開催するなど活発に活動を続けている。

研究成果の概要(英文)：Compilation of the 50-million-word balanced corpus of contemporary Japanese books was successfully achieved according to the initial plan. The corpus was publicly available since 2011. Various studies using the corpus under compilation have been conducted to achieve the goal of establishing the basis of corpus Japanese linguistics. Research workshops consisting of more than 200 researchers have been held regularly even after the end of the current project.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	9,600,000	0	9,600,000
２００７年度	11,700,000	0	11,700,000
２００８年度	11,600,000	0	11,600,000
２００９年度	11,600,000	0	11,600,000
２０１０年度	19,700,000	0	19,700,000
総　計	64,200,000	0	64,200,000

研究分野：人文学/総合領域
科研費の分科・細目：言語学・日本語学/情報学・知能情報学
キーワード：均衡コーパス、日本語、書き言葉、代表性

１．研究開始当初の背景

言語コーパスは 21 世紀の言語研究にとって欠くことのできないインフラであるが、日本語コーパスの整備状況は、英米はもとより、韓国・台湾などアジア諸国に比べても出遅れており、コーパス日本語学を推進するためのインフラが整っていない状態であった。

２．研究の目的

第一の目標は 5000 語規模の現代日本語書籍の均衡コーパスを構築することであった。第二の目標はコーパスの構築と並行して、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開拓することであった。

３．研究の方法

コーパスの構築に関しては、国立国語研究所が中心となり、同研究所の KOTONOHA 計画と連携して、合計で 1 億語を超す『現代日本語書き言葉均衡コーパス』を構築することを目標とした。また国語研と千葉大学が連携し

て形態素解析用辞書 UniDic を開発することとした。さらに奈良先端大が中心となってコーパスの利用環境の整備を進めることとした。

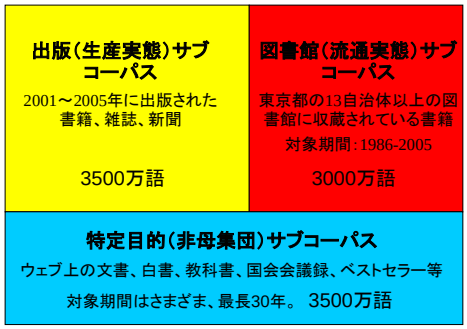
コーパス日本語学の開拓については、日本語学、自然言語処理、日本語教育学、国語教育学、辞書学などの関連領域に計画研究班を設置した。基礎研究のみならず、応用研究を同時に推進することとした。

また第 2 年次からは公募班による研究も推進した。

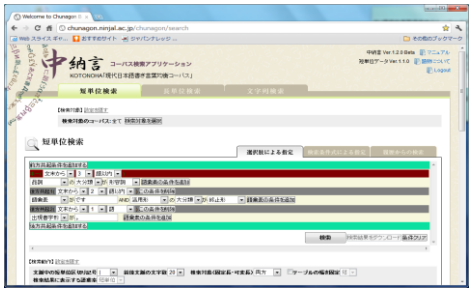
４．研究成果

(1) コーパスの公開

『現代日本語書き言葉均衡コーパス』（以下 BCCWJ）に含まれる書籍データ（6000 万語以上）は、本領域で構築したものである。図に BCCWJ の内部構造を示す。書籍コーパスは、出版サブコーパスのうち 3000 万語と図書館サブコーパスの全体（3000 万語）および特定目的サブコーパスのうちベストセラーが該当する。



BCCWJ の内部構造



『中納言』の検索画面

BCCWJ のサンプルには著作権処理が施されており一般公開が可能である。2011 年 3 月には全文検索用検索ウェブサイトである『少納言』による無償公開を開始した。現在までに延べで 4 万名を超える利用者がある。

その後同年 8 月には形態論情報を検索するためのウェブインターフェース『中納言』を公開し、12 月にはデータ全体を DVD 版として公開することによって公開作業を終了した。

これまで『中納言』には 400 件以上の、また DVD 版には 100 件以上の利用申請があった。

(2) 形態論情報解析用辞書の公開

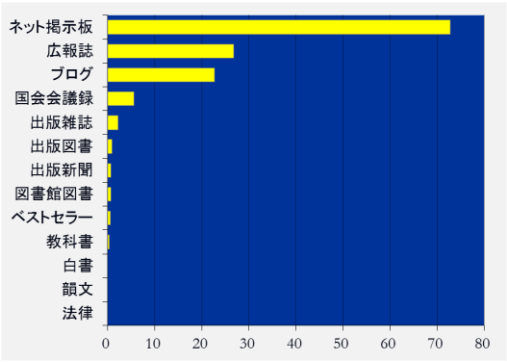
『中納言』で公開している形態論情報は、BCCWJ に格納されている多種多様なテキストジャンル（新聞、白書、書籍、雑誌、ブログ、ネット掲示板、国会会議録等、広報誌、法律）のすべてにおいて、目標とした 98%の解析精度を達成した。

そのために整備した解析用辞書 UniDic も一般公開した。ダウンロード回数は 3500 回以上に及んでいる。民間企業によってパソコンやスマートホンの OS にも搭載されている。

(3) 日本語の運用実態の解明

本領域の活動期間のみならず、領域の終了後も多くの研究者によって、現代日本語の運用実態に関する調査が、BCCWJ を利用して実施されている。ここでは文法と文字に関する分析例を紹介する。

下図は「白いです。」「早いです。」のように形容詞に助動詞「です」が後続して文が終わるタイプの述語の生起率(%)を『中納言』を用いて検索し、BCCWJ のジャンル別に示したものである。この種の形容詞述語は日本語文法では容認性が低いとされているが、ジャンルによっては 7 割以上の高い生起率を示しており、一概に容認性が低いと判断することには無理があることが分かる。



「形容詞＋です。」の生起率(%)

BCCWJ は漢字の運用についても貴重な情報を提供する。構築途上の出版サブコーパス中の書籍サンプルに含まれる文字量を調査したところ、延べで 21,379,644 字、異なりで 6,071 字であったが、JIS X 0213 で表現できない JIS 外字は異なりで 142 字、延べで 284 字(0.001%)に過ぎないことが分かった。

いわゆる JIS 第 3，第 4 水準までの漢字があれば現代日本語の外字問題は（細かな字体の問題を無視すれば）ほとんど解決することを示す結果であるが、このような調査結果は従来報告されたことがなかった。

(4)言語教育への応用

BCCWJ の応用が最も期待される領域が言語教育（国語教育及び日本語教育）である。外国語教育におけるコーパスの重要性は、英語教育等において早くから知られているが、日本人を対象とした国語教育における基礎語彙の選定等もコーパスを利用することによって、客観的な基準によって実施することが可能になる。

例えば、所与の語の頻度順位を BCCWJ の様々なジャンル間で比較することで、語の性格を探ることができる。公的なジャンル（書籍や新聞）では順位が高いが、私的なジャンル（ネット掲示板やブログ）では順位が低い語には、以下のようなものがある（五十音順に上位のみを示す）。

語種	語彙
和語	息衝く、憤り、戒める、浮き上がる、受け皿、移り住む、裏付ける、推し進める、躍り出る…
漢語	医科、遺構、異質、移送、一元、一座、一助、一団、一門、一室、一世、異物、院内、右派…
外来語	エコロジー、カリキュラム、ジャンボ、ナショナリズム、ノンフィクション、ハイジャック…

反対に私的なジャンルにおける順位が高いものは以下のような語である。

語種	語彙
和語	荒らし、うざい、腕立て、大泣き、起きる、男前、着る、出し、立ち読み、食べる…
漢語	完済、視聴、洗車、駐輪、年式、納車、模試、陸運…
外来語	アウトレット、アド、アプリ、エアロ、エディション、オムライス、カラーリング…

このような語の性格付けは教育用語彙の選択を科学的に実施する可能性を開く出発点である。

(5)BCCWJ への拡張アノテーション

現在の BCCWJ には、形態論情報のほかに、著者情報、書誌情報、文章構造（章、段落、節、文、図表等の情報）の情報が提供されている。しかし研究目的によっては、より多くの研究用付加情報（アノテーション）が必要とされる。そこで BCCWJ 全体から 100 万語規模のサブコーパスである「コア」を選定し、コアに対して種々のアノテーションを施した。文節係り受け構造、多義語の語義分類、動詞項構造、フレームネット構造などのアノテーションを本領域実施期間中に実施した。一部の作業は領域終了後も継続して実施されている。

(6)研究成果の普及

本領域の成果を広く普及させるために 8 巻構成の『講座日本語コーパス』（朝倉書店）の刊行を準備しており、順調に進めば平成 24 年度中に刊行が始まる予定である。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 416 件）

1) Kikuo Maekawa. “Coarticulatory reinterpretation of allophonic variation: Corpus-based analysis of /z/ in spontaneous Japanese.” *Journal of Phonetics*, 査読有, 38(3), 2010, pp.360-374.
2) Kikuo Maekawa, Yamazaki, M., Maruyama, T., Yamaguchi, M., Ogura, H., Kashino, W., Ogiso, T., Koiso, H., & Den, Y . “Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese.” *Proc. 7th International Conference on Language Resources and Evaluation* (LREC2010), 査読有, 2010, pp. 1483-1486.
3) 飯田龍・小町守・井之上直也・乾健太郎・松本裕治.「述語項構造と照応関係のアノテーション：NAISTテキストコーパス構築の経験から」, 自然言語処理, 査読有, Vol.17, No.2, 2010, pp.25-50.
4) 田野村忠温.「日本語コーパスとコロケーション―辞書記述への応用の可能性―」, 『言語研究』, 査読有, 第138号, 2010, pp.1-23.
5) Kyoko Hirose Ohara. “Frame-based contrastive lexical semantics in Japanese FrameNet: The case of risk and kakeru.” In Hans C. Boas (Ed.) *Multilingual FrameNets in Computational Lexicography:Methods and Applications*, Mouton de Gruyter, 査読有, 2009, pp.163-182.
6) 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, 査読有, 4(1), 2008, pp.82-95.
7) 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, 査読有, 15, 2008, pp.3-16.
8) 岩立将和・浅原正幸・松本裕治.「トーナメントモデルを用いた日本語係り受け解析」, 自然言語処理, 査読有, Vol.15, No.5, 2008, pp.169-185.
9) Ryu Iida, Kentaro Inui, Yuji Matsumoto. “Zero-anaphora resolution by learning rich syntactic pattern features.”, *ACM Transactions on Asian Language Information Processing* (TALIP), 査読有, Vol 6, Issue 4, 2007, pp.1-22.

10) Irena Srdanovic Erjavec, Andrej Bekes, Kikuko Nishina.“ Cluster analysis of suppositional adverbs and clause-final modality.”, *Asian and African Studies*, University of Ljubljana Faculty of Arts, 査読有, Vol. XI, No.3, 2007, pp. 21-31.
11) 小椋秀樹・相澤正夫.「現代雑誌70誌における漢字の使用実態と常用漢字表―国語施策へのコーパス活用に向けた基礎調査―」日本語科学, 査読有, 22, 2007, pp.125-146.
12) 伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵.「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」, 日本語科学, 査読有, 22, 2007, pp. 101-122.
13) 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―」, 査読有, 日本語科学, 22, 2007, pp.13-28.
14) Chu-Ren Huang, Takenobu Tokunaga, Sohpia Yat Mei Lee. “Asian language processing: current state-of-the-art.”, *Language Resources and Evaluation*, 査読有, Vol.40, No.3-4, 2006, pp.203-218.
15) 荻野綱男.「形容動詞連体形における「な／の」選択について―田野村氏の結果をWWW で調べる―」計量国語学, 査読有, Vol.25, No.7, 2006, pp.309-318.

〔学会発表〕（計 542 件）
1) 服部匡「国会会議録データと現代語の通時変化」,日本語学会2010年度秋季大会ワークショップ「コーパス日本語学の新展開―コーパスと方法論の多様化―」, 2010年10月23日, 愛知大学.
2) Kikuo Maekawa. “KOTONOHA: A Corpus Compilation Initiative at the National Institute for Japanese Language.”, Keynote speech at the 22nd International Conference on the Computer Processing of Oriental Languages, March 27, 2009, Hong Kong Polytechnic University, Hong Kong.
3) Kikuo Maekawa. “Compilation of the Balanced Corpus of Contemporary Written Japanese in the KOTONOHA Initiative.” Second International Symposium on Universal Communication, December 15, 2008, Osaka International Convention Center.
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―」, 2008年12月1日, 学習院創立百周年記念会館.
5) 小磯花絵・小木曽智信・小椋秀樹・冨士池優美・宮内佐夜香.「『現代日本語書き言葉均衡コーパス』にもとづくジャンル間の文体差に関わる要因の分析」, 第22回社会言語科学会研究大会, 2008年9月13日, 愛知大学.
6) 小原京子.「コーパスに基づく日本語主観移動表現のフレーム意味論的分析：英語との比較から」, 日本認知科学会第25回大会, 2008年9月5日, 同志社大学京田辺校地.
7) 砂川有里子.「日本語教育におけるコーパス活用の可能性」, 日本語教育世界大会2008, 2008年7月, 釜山外国語大学校, 韓国.
8) 藤井聖子・上垣渉.「支援動詞構文における事態性名詞と動詞との項共有と連結性:『日本語コーパス』を用いた分析」, 日本言語学会第136回大会, 2008年6月22日, 学習院大学.
9) 相澤正夫・小椋秀樹.「白書コーパスに基づく常用漢字の使用実態調査」, 第21回社会言語科学会研究大会, 2008年3月22日, 東京女子大学.
10) 田中牧郎・金愛蘭・桐生りか・近藤明日子.「コーパスによる難解語・重要語の抽出―医療用語を例に―」, 社会言語科学会第21回大会, 2008年3月22日, 東京女子大学.
11) 佐野大樹・丸山岳彦.「システミック文法に基づく書きことばの複雑さ測定―日本語大規模コーパスを用いた語彙密度計測―」, 言語処理学会第14回年次大会, 2008年3月20日, 東京大学駒場キャンパス.
12) 冨士池優美・小椋秀樹・小木曽智信・小磯花絵・内元清貴・相馬さつき・中村壮範.「『現代日本語書き言葉均衡コーパス』の長単位認定基準について」, 言語処理学会第14回年次大会, 2008年3月20日, 東京大学駒場キャンパス.
13) 小木曽智信・小椋秀樹・伝康晴.「日本語研究に適した形態素解析ソフトウェア―「unidic」と「茶まめ」―」, 日本語学会2007年度秋季大会, 2007年11月18日, 沖縄国際大学.
14) 柏野和佳子.「国語辞典における多義語の意味記述の比較」, 言語処理学会第13回年次大会, 2007年3月21日, 龍谷大学.
15) 飯田龍・小町守・乾健太郎・松本裕治.「NAISTテキストコーパス: 述語項構造と共参照関係のアノテーション」, 情報処理学会第177回自然言語処理研究会, 2007年1月26日, 筑波大学春日キャンパス.

〔図書〕（計 10 件）

1) Kyoko Hirose Ohara and Nikiforidou, Kiki (Eds.), John Benjamins Publishing Company, *Constructions and Frames.*, 2010.
2) 砂川有里子他, くろしお出版, 『日本語教育研究への招待』, pp.99-119および pp.193-211, 2010.
3) 山内博之, ひつじ書房, 『プロフィシェンシーから見た日本語教育文法』, 185p, 2009.
4) 山内博之編著, ひつじ書房, 『日本語教育スタンダード試案 語彙』, 120p, 2008.
5) Yukio Tono, Peter Lang Pub Inc., *Spoken Corpora in Applied Linguistics*. 264p, 2007.
6) 柴崎秀子, 風間書房, 『第二言語テキスト理解と読み手の知識』, 181p, 2006.

〔その他〕
ホームページ等
①http://www.tokuteicorpus.jp/
②http://www.kotonoha.gr.jp/demo/
③http://www.kotonoha.gr.jp/shonagon/
④https://chunagon.ninjal.ac.jp
⑤http://www2.ninjal.ac.jp/kikuo/

６．研究組織
(1)研究代表者
前川　喜久雄（MAEKAWA KIKUO）
国立国語研究所・言語資源研究系・教授
研究者番号：20173693

(2)研究分担者
山崎　誠（YAMAZAKI MAKOTO）
国立国語研究所・言語資源研究系・准教授
研究者番号：30182489
松本　裕治（MATSUMOTO YUJI）
奈良先端科学技術大学院大学・ 情報科学研究科・教授
研究者番号：10211575
傳　康晴（DEN YASUHARU）
千葉大学・文学部・教授
研究者番号：70291458
田野村　忠温（TANOMURA TADAHARU）
大阪大学･大学院文学研究科・教授
研究者番号：40207204
砂川　有里子（SUNAKAWA YURIKO）
筑波大学・人文社会科学研究科・教授
研究者番号：40179289
田中　牧郎（TANAKA MAKIRO）
国立国語研究所・言語資源研究系・准教授
研究者番号：90217076
荻野　綱男（OGINO　TSUNAO）
日本大学・文理学部･教授
研究者番号：00111443
奥村　学（OKUMURA　MANABU）
東京工業大学・精密工学研究所・教授
研究者番号：60214079
斎藤　博昭（SAITO HIROAKI）
慶應義塾大学・理工学部・准教授
研究者番号：30235064
（H.19→H.20）
柴崎　秀子（SHIBASAKI HIDEKO）
長岡技術科学大学・工学部・教授
研究者番号：00376815
（H.19→H.20）
新納　浩幸（SHINNO HIROYUKI）
茨城大学・工学部・准教授
研究者番号：10250987
（H.19→H.20）
仁科　喜久子（NISHINA KIKUKO）
東京工業大学・留学生センター・ 教授
研究者番号：40198479
（H.19→H.22）
宇津呂　武仁（UTSURO TAKEHITO）
筑波大学・システム情報工学研究科・准教授
研究者番号：90263433
（H.21→H.22）
関　洋平（SEKI YOHEI）
筑波大学・大学院図書館情報メディア研究科・助教
研究者番号：00348468
（H.21→H.22）
小原　京子（OHARA KYOKO）
慶應義塾大学・理工学部・准教授
研究者番号：00286650
（H.21→H.22）

(3)連携研究者
木戸 冬子（KIDO FUYUKO）
東京大学・大学院情報理工学系研究科・特任助教
研究者番号：60527828