様式 C-19

科学研究費補助金研究成果報告書

平成 23 年　6 月　16 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：１８０６１００７
研究課題名（和文）　代表性を有する現代日本語書籍コーパスの構築

研究課題名（英文）　Compilation of a balanced book corpus of contemporary written Japanese

研究代表者
山崎　誠（YAMAZAKI MAKOTO）
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・准教授
研究者番号：30182489

研究成果の概要（和文）：本研究では、今後の日本語研究にとって利用価値の高い、大規模な書籍コーパスを構築した。この書籍コーパスは、以下の特徴を持つ日本で最初の本格的な書き言葉コーパスである。(１)ランダムサンプリングによって母集団を過不足なく代表すること。(２)短単位・長単位の２種類の言語単位による形態論情報が付与されていること。(３) XML により文書構造、形態論情報、文字情報等が記述されていること。(４) 可能な限りすべてのサンプルの著者に連絡をとり著作権処理を施したこと。書籍コーパスは『現代日本語書き言葉均衡コーパス』の主要な部分を占め、２０１１年中に一般に公開する。

研究成果の概要（英文）：We have compiled a large balanced corpus of books which will be a highly useful resource for the future research of Japanese language. This corpus is the first authentic balanced written corpus in Japan and has the following characteristics.(1)Represents the distribution of population properly by random sampling. (2)Segmented by two kinds of word unit(short word unit and long word unit). (3)Text strucrure, morphological information and character information are annotated using XML.(4)Every sample is sought the copyright permission as long as possible. The book corpus is the main part of the BCCWJ(Balanced Corpus of Contemporary Written Japanese) and will be open to the public in 2011.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	54,300,000	0	54,300,000
２００７年度	86,200,000	0	86,200,000
２００８年度	64,900,000	0	64,900,000
２００９年度	33,408,000	0	33,408,000
２０１０年度	17,500,000	0	17,500,000
総　計	256,308,000	0	256,308,000

研究分野：人文学
科研費の分科・細目：言語学・日本語学
キーワード：均衡コーパス，書き言葉，代表性，書籍，サンプリング，XML，形態解析，著作権処理

１．研究開始当初の背景

１９９０年代におけるコンピュータ環境の普及に伴い、「コーパス」を利用した言語研究への関心が高まってきた。コーパスを活用した言語研究で先んじていたのは英語を対照とした英語コーパス言語学の世界であった。一方、日本語研究においては、工学系の言語処理研究において実用的な目的でコーパスの利用が積極的に進められていた。それに比して人文系の日本語研究においては、

コーパスそのものの不足から研究のブレイクスルーが見いだせない状況であった。海外では、英語の代表的なコーパスである British National Corpus の完成（１９９４年）に触発され、諸外国でも国家プロジェクトとしてコーパスの構築が始まっていた。そのような中で、日本でも国立国語研究所から２００４年に『日本語話し言葉コーパス』、２００５年に『太陽コーパス』が公開され、コーパスが研究に与える飛躍的貢献が認められるようになり、当時の日本語研究でもっとも需要が高かった現代語の書き言葉のコーパス作成への期待が学界内に高まってきた。世界的な水準の書き言葉コーパスを効率良く達成するには、既に二つのコーパスを公開して、コーパス構築のノウハウを備えた国立国語研究所がもっとも適切であることから、特定領域研究を申請するに至った。

２．研究の目的

本研究の目的は、日本におけるコーパス言語学の定着と発展及びコーパスを利用した応用技術の深化と拡大を目指して、書籍を対象にそれらを適切に代表する「書籍コーパス」の構築方法を開発し、その方法論に基づき、実際にコーパスを完成させることである。

３．研究の方法

(1)書籍コーパスの代表性を確保するために、書き言葉の生産と流通という２つの観点から母集団を設定する。生産の面では、２００１年～２００５年に国内で出版された書籍を対象に、流通の面では、東京都内の５２自治体の公立図書館に所蔵されている書籍のうち、１９８６年～２００５年に発行され、かつ、１３自治体以上で所蔵されている書籍を対象とした。

(2)書籍コーパスの構築作業はサンプリング、著作権処理、電子化、形態論情報付与の４つの段階に分かれる。（１）サンプリングでは、無作為抽出により、固定長サンプル（１００００字）と可変長サンプル（１万語を上限とするひとまとまりの文章）を指定し、併せて、書誌情報を取得する。（２）著作権処理では、書誌情報をもとに抽出されたサンプルの著作者を推定し、許諾依頼を行う。（３）電子化では、ＸＭＬ（文書構造記述言語）を用いて分析用の様々な情報を付ける。（４）形態素解析では、自動的に精度の高い単語認定を行う。目標とする精度は、９８％以上である。

(3)コーパスの一部を「コアデータ」として、人手による修正を加え、精度の高いデータとして活用できるようにする。

(4)コーパスの諸仕様は、マニュアルの形で整備して公開し、コーパス利用の普及をはかる。

構築途中のコーパスを「モニター公開データ」として希望者に無料で配布し、利用状況等を把握し、構築のための参考情報を得る。

４．研究成果

(1)書籍コーパス（24,320 サンプル、約 6500 万語）の構築をほぼ完了した。書籍コーパスは『現代日本語書き言葉均衡コーパス』の主要な部分を占めており、その他のサブコーパスと併せて一般公開の準備を進める段階に入った。書籍コーパスは、すべてのサンプルに対して固定長サンプル、可変長サンプルを設け、分析目的に応じて使い分けられるようにした。

書籍コーパスを構成する生産実態サブコーパス及び流通実態サブコーパスについて，NDC(日本十進分類法)別の構成比を示す。比率は固定長サンプルで計測したものである。

表１　出版実態サブコーパスの構成比

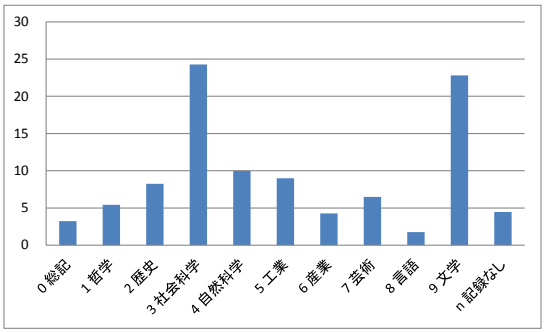
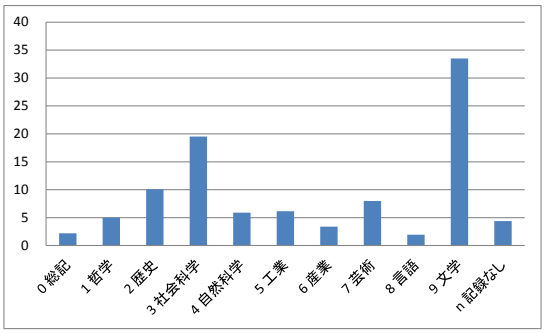


表２　流通実態サブコーパスの構成比



(2)形態素解析用辞書 UniDic に搭載する短単位データベースを電子化辞書班と連携して構築した。辞書の見出しに当たる語彙素数は約２１万である。

(3)形態素解析結果を人手で修正し精度を９９％に高めたコアデータを完成させた。コアデータには、短単位、長単位、文節の３種類の言語単位の境界情報を付与した。語数は短単位で約１１０万である。

コアデータの内訳は以下の表３のとおり。

表３　コアデータの内訳

媒体	短単位	長単位
雑誌	246,000	200,000
新聞	361,000	273,000
白書	228,000	159,000
Yahoo!知恵袋	111,000	95,000
Yahoo!ブログ	118,000	100,000

(4)著作権処理はすべての書籍サンプルについて連絡先を調査し、連絡先が判明したものについては許諾依頼を行った。
(5)コーパスの構築に関するノウハウをまとめた報告書を１７冊刊行した。

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３６件）
(1)佐野大樹，小磯花絵(2011)，「現代日本語書き言葉における修辞ユニット分析の適用性の検証－『書き言葉らしさ・話し言葉らしさ』と脱文脈化言語・文脈化言語の関係－」，機能言語学研究，6，ページ数不明，査読有
(2)佐野大樹(2010)．「『話し言葉らしさ・書き言葉らしさ』の計測－語彙密度の日本語への適用性の検証」機能言語研究，5，pp.89-102，査読有
(3)佐野大樹(2010)．「ブログにおける評価表現の使い分けの特徴―アプレイザル理論からみた評価基準と表現の直接性/間接性の関係―」計量国語学，27(7)，pp.249-269，査読有
(4)Wakako Kashino，Manabu Okumura(2010)．”An Approach toward Register Classification of Book Samples in the Balanced Corpus of Contemporary Written Japanese”，24th Pacific Asia Conference on Language，Information and Computation (PACLIC24)，2010.11.6，東北大学，査読有
(5)山崎誠(2009)．「代表性を有する大規模日本語書き言葉コーパスの構築」人工知能学会誌，24(5)，pp.623-631，不詳
(6)丸山岳彦(2008)．「『日本語話し言葉コーパス』に基づく言い直し表現の機能的分析」，『日本語文法』8巻2号，pp.121-139，査読有
(7)Sano Motoki，Takehiko Maruyama(2008)．″Lexical Density in Japanese Texts: classifying text samples in Balanced Corpus of Contemporary Written Japanese″，Proceedings of 35th International Systemic Functional Congress，pp.359-364，2008.7.22，Sydney.，査読有
(8)Takehiko Maruyama，Makoto Yamazaki，Kikuo Maekawa(2008)．″Statistical sampling method used in the Balanced Corpus of Contemporary Written Japanese″，18th International Congress of Linguists，2008.7.25，Seoul.，査読有
(9)小椋秀樹(2008)．「コーパスのための形態論情報」，『国文学解釈と鑑賞』74-1，pp.26-34，至文堂，不詳
(10)山崎誠(2008)．「国立国語研究所における諸研究－語彙調査の系譜の中心にして－」，『国文学解釈と鑑賞』74-1，pp.183-191，至文堂，不詳
(11)丸山岳彦，田野村忠温(2007)．「コーパス言語学の射程」『日本語科学』22，pp.5-12，2件

〔学会発表〕（計１０４件）
(1)冨士池優美，小西光，小椋秀樹ほか(2011)「長単位に基づく『現代日本語書き言葉均衡コーパス』の品詞比率に関する分析」－BCCWJの文書構造情報分析を中心に－」，言語処理学会第17回年次大会，2011年3月9日，豊橋技術科学大学
(2)柏野和佳子，奥村学(2010)．「国語辞典に「古い」と注記される語の現代書き言葉における使用傾向の調査」，第88回人文科学とコンピュータ研究会発表会，2010年10月30日，国立国語研究所．
(3)山崎誠(2010)．「テキストにおける多義語の意味実現の傾向」，計量国語学会第54回大会横言う集pp.25-30，2010年9月11日，大正大学．
(4)柏野和佳子(2010)．「「直接的な語り」という表現スタイルをもつ書籍テキストの人手抽出の試み」，第 35 回ことば工学研究会，2010年8月28日，神奈川大学横浜キャンパス．
(5)Maekawa，Kikuo，Makoto Yamazaki，Takehiko Maruyama，Masaya Yamaguchi，Hideki Ogura，Wakako Kashino，Toshinobu Ogiso，Hanae Koiso and Yasuharu Den(2010)．”Design，Compilation，and Preliminary Analyses of Balanced Corpus of Contemporary Written Japanese”，7th International Conference on Language Resources and Evaluation (LREC2010) ，2010.5.20，Mediterranean conference centre，Valleta，Malta．
(6)間淵洋子(2010)．「コーパスを用いたテキスト分類指標の検討－BCCWJの文書構造情報分析を中心に－」，言語処理学会第16回年次大会，2010年3月10日，東京大学．
(7)小椋秀樹(2010)．「形態素解析辞書UniDicにおける同語異語判別について」，言語処理学会第16回年次大会，2010年3月10日，東京大学．
(8)丸山岳彦(2010)．「代表性を有するコーパスの設計とサンプリングの実際－コーパスに基づく言語研究の可能性と限界－」，言語

処理学会第16回年次大会,2010年3月9日,
東京大学.
(9)柏野和佳子(2009).「和語や漢語のカタカナ表記:『現代日本語書き言葉均衡コーパス』における使用実態」,計量国語学会第54回大会予稿集,pp.38-43,2009年9月12日,東京女子大学.
(10)山崎誠,丸山岳彦,柏野和佳子,佐野大樹,秋元祐哉,稲益佐知子,田中弥生,大矢内夢子(2009).「現代日本語書き言葉均衡コーパスのサンプル長と言語的特徴-固定長サンプルと可変長サンプルの質的な違い-」言語処理学会第15回年次大会[NLP2009]予稿集pp.618-621,
2009年3月5日,鳥取大学.
(11)柏野和佳子,丸山岳彦,稲益佐知子,秋元祐哉,田中弥生,佐野大樹,大矢内夢子,山崎誠(2009).「『現代日本語書き言葉均衡コーパス』のサンプル収録方法」言語処理学会第15回年次大会[NLP2009]予稿集pp.196-199,2009年3月3日,鳥取大学.
(12)柏野和佳子(2008).「書籍の文章の多様性をとらえる観点付与の設計―『現代日本語書き言葉均衡コーパス』の収録文章を対象に―」第30回ことば工学研究会予稿集pp.11-22,
2008年10月31日,武蔵野美術大学新宿サテライトROOM.
(13)佐野大樹(2008).「大規模バランストコーパスにおけるテクスト分類に向けて-語彙密計測からみたコンテクスト情報-」日本機能言語学会第16回秋季大会,2008年10月12日,お茶の水女子大学.
(14)佐野大樹,丸山岳彦(2008).「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[NLP2008]予稿集,pp.1097-1100,2008年3月20日,東京大学.
(15)柏野和佳子,丸山岳彦,秋元祐哉,稲益佐知子,佐野大樹,田中弥生,山崎誠(2008).「書籍の生産実態を反映するサンプリング―NDCごとに取得したサンプルの多様性の分析―」言語処理学会第14回年次大会[NLP2008]予稿集,pp.939-942,2008年3月20日,東京大学.
(16)小椋秀樹,小木曽智信,原裕,小磯花絵,冨士池優美(2008).「形態素解析用辞書UniDicへの語種情報の実装と政府刊行白書の語種比率の分析」言語処理学会第14回年次大会[NLP2008]予稿集,pp.935-938,2008年3月20日,東京大学.
(17)山崎誠,丸山岳彦,山口昌也,小椋秀樹,森本祥子,柏野和佳子,佐野大樹,高田智和,間淵洋子,北村雅則,小木曽智信,小磯花絵,冨士池優美,小沼悦,田中牧郎,前川喜久雄(2007).「現代日本語書き言葉均衡コーパスの設計と検索デモンストレーション」日本語学会2007年度秋季大会(沖縄国際大学)予稿集,pp.239-246,2007年11月18日,沖縄国際大学.
(18)丸山岳彦,柏野和佳子,稲益佐知子,秋元祐哉,吉田谷幸宏,山崎誠(2007).「書き言葉の構造を捉える―書き言葉の多様な構造とサンプリング手法―」言語処理学会第13回年次大会[NLP2007]予稿集pp.704-707,
2007年3月21日,龍谷大学.
(19)小椋秀樹,小木曽智信,小磯花絵,冨士池優美,相馬さつき(2007).「「現代日本語書き言葉均衡コーパス」の短単位解析について」言語処理学会第13回年次大会[NLP2007]予稿集pp.720-723,2007年3月21日,龍谷大学.
(20)高田智和,山口昌也(2006).「文字・表記研究とコーパス」漢字文献情報処理研究会第9回大会,2006年12月16日,ピアザ淡海305会議室(大津市).
(21)山崎誠,丸山岳彦,柏野和佳子,前川喜久雄,稲益佐知子,秋元祐哉,吉田谷幸宏(2006).「現代日本語書き言葉均衡コーパスのサンプリング方法について」計量国語学会第50回大会,
2006年9月30日,国立国語研究所.

〔図書〕(計　件)

〔その他〕
ホームページ等
・特定領域「日本語コーパス」
http://www.tokuteicorpus.jp/
・KOTONOHA『現代日本語書き言葉均衡コーパス』検索デモンストレーション
http://www.kotonoha.gr.jp/demo/

６．研究組織
(1)研究代表者
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
(2)研究分担者
丸山 岳彦(MARUYAMA TAKEHIKO)
国立国語研究所・言語資源研究系・助教
研究者番号:90392539
柏野 和佳子(KASHINO WAKAKO)
国立国語研究所・言語資源研究系・准教授
研究者番号:50311147
佐野 大樹(SANO MOTOKI)
国立国語研究所・コーパス開発センター・プロジェクト特別研究員
研究者番号:60455425
(H20→H22)

山口 昌也（YAMAGUCHI MASAYA）
国立国語研究所・言語資源研究系・助教
研究者番号：30302920
間淵 洋子（MABUCHI YOKO）
国立国語研究所・コーパス開発センター・プロジェクト特別研究員
研究者番号：10415614
高田 智和（TAKADA TOMOKAZU）
国立国語研究所・理論・構造研究系・准教授研究者番号：90415612
小椋 秀樹（OGURA HIDEKI）
国立国語研究所・言語資源研究系・准教授
研究者番号：00321547
冨士池 優美（FUJIIKE YUMI）
国立国語研究所・コーパス開発センター・プロジェクト特別研究員
研究者番号：20510572
（H20→H22）
小沼 悦（ONUMA ETSU）
国立国語研究所・管理部研究推進課・専門職員
研究者番号：00311150
（H19→H22）
森本 祥子（MORIMOTO SACHIKO）
学習院大学大学院・人文科学研究科・助教
研究者番号：80342939
（H18→H20）
大和 淳
文化庁長官官房著作権課・課長補佐
研究者番号：10377103
（H18）