様式 C-7-2

自己評価報告書

平成 21 年 5 月 28 日現在

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）
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：30182489

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．研究計画の概要

本研究では，現代日本語研究にとってもっとも重要な研究基盤と位置付けられる「書籍コーパス」を構築することを目的とする。このコーパスは，従来の新聞や文学作品といった単一のジャンルから構成される電子資料と違って，多様なジャンルや文体を持つ書籍を資料として，その資料的あるいは言語的諸特性を適切に代表するコーパスとして設計する。そのために，ランダムサンプリング，ＸＭＬによるタグに記述，斉一的な言語単位による形態素情報の付与，著作家処理を行い，多方面での高度な活用を目指す。

２．研究の進捗状況

○2006 年度

コーパスの全体設計及び電子化仕様の確定，サンプリングにおける近隣図書館からの協力・著作権処理における作家団体からの協力の獲得など、構築体制の基盤を固めた。約 600 万語（書籍）の構築し，報告書 2 冊刊行した。

○2007 年度

コーパスの規模を約 3400 万語（書籍）に増やし，書籍全体の 50%をサンプリングと入力を完了した。コアデータの設計と構築，形態素解析用辞書の整備等，付加情報の充実を図った。検索デモンストレーションサイトの公開(5 月)し，著作権処理の円滑化を目指した。報告書 4 冊を刊行した。

○2008 年度

書籍全体の約 80%のサンプリングと入力を完了し，コーパスの規模を約 5200 万語（書籍）に増やした。コアデータ 50 万語分を領域内公開するとともに，長単位による解析も開始した。モニター公開データ(2800 万語)の配布を開始(7 月)。報告を 3 冊刊行した。

３．現在までの達成度

①当初の計画以上に進展している。

2008 年度末で必要書籍サンプル数 25208 に対してサンプリング 78.9%，電子化（入力）77%，ＸＭＬ化 73%を達成した。ＸＭＬ化まで終了した語数は推定 5400 万語で，計画の目標である 5000 万語を突破した。

４．今後の研究の推進方策

2009 年度：残りの約 20%の書籍の入手，サンプリング，電子化，著作権処理に努め，同時に形態素解析の精度向上のため，短単位データベースへの未知語登録を進める。

2010 年度：サンプルのジャンル別構成比に留意しながらサンプルの取得を進め，予定どおりの均衡コーパスを完成させる。

５．　代表的な研究成果

（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 24 件）

(1)小椋秀樹「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，2008:12,査読無
(2)山崎誠「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，2008:12,査読無
(3)佐野大樹「学術的表現への言い換え―教育現場での選択体系機能言語理論―」『日本語学』26(13)，pp.60-71，2007:11,査読無
(4)丸山岳彦・田野村忠温「コーパス言語学

の射程」『日本語科学』22, pp. 5-12, 2007:10, 査読無
(5)柏野和佳子「書き言葉コーパスで探る日本語のありさま」, 日本語学 25-9, pp. 18-27, 2006:08, 査読無
〔学会発表〕(計 77 件)
(1)Maruyama, Takehiko, Makoto Yamazaki, and Kikuo Maekawa, ″Statistical sampling method used in the Balanced Corpus of Contemporary Written Japanese″, 18th International Congress of Linguists. Seoul, 2008:07
(2)小椋秀樹, 小木曽智信, 原裕, 小磯花絵, 冨士池優美「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](東京大学)予稿集, pp.935-938, 2008:03
(3)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「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](東京大学)予稿集, pp.939-942, 2008:03
(4)高田智和・間淵洋子・西部みちる・北村雅則・山口昌也「文字コードとタグによる漢字字体の記述」言語処理学会第 13 回年次大会 [NLP2007](龍谷大学)予稿集 pp.712-715,2007:03
(5)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「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](龍谷大学)予稿集 pp.704-707,2007:03
(6)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「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(国立国語研究所), 2006:09

〔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〕