様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ５ 月 ３１ 日現在

機関番号：１４４０１
研究種目：挑戦的萌芽研究
研究期間：2013～2015
課題番号：２５５４０１６３

研究課題名（和文）XMLコーパスからの抽出データに基づく日本語学術ライティング教材作成法の研究

研究課題名（英文）Research on Developing Learning Materials for the Japanese Academic Writing Based on the Mined Data from XML Corpora

研究代表者
堀　一成（Hori, Kazunari）

大阪大学・全学教育推進機構・准教授

研究者番号：８０２７０３４６

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：学術的な文章の言語資源（コーパス）から言語特徴抽出を行い、ライティング教育の学習コンテンツとして提供することを目的に推進し、本研究においては以下の３点の主な成果を得た。
(1) XMLコーパスである現代日本語書き言葉均衡コーパス(BCCWJ)の高校教科書データを対象に解析し、用いられている動詞・名詞の頻度情報を得た。(2)平文テキストデータに長単位形態素情報を付与するためのソフトウェアComainuを利用し、学術文・技術文データの処理手順の開発と処理作業を行った。(3)学術文(約6,000字)、技術文(約15,000字)から情報抽出し、長単位動詞・名詞の頻度データを得た。

研究成果の概要（英文）：We have mined linguistic characteristics from academic corpora, in order to provide them as learning materials for our writing classes, and so far obtained the following three results.
(1)Obtained the frequency of verbs and nouns used in the high school textbook data from “The Balanced Corpus of Contemporary Written Japanese”(BCCWJ).(2)Developed and executed the procedure for mining both academic and technical sentences, using Comainu software which imparts the long-unit-word morphological information to plain text data.(3)Obtained the long-unit-word verb and noun frequency through mining information from academic sentences (approx. 6,000 characters) and technical sentences (approx. 15,000 characters).

研究分野：自然言語処理、数理工学

キーワード：アカデミック・ライティング　データマイニング　日本語コーパス　大学リポジトリ　学習コンテンツ　コロケーション　BCCWJ　eラーニング

1. 研究開始当初の背景

論文・レポートの書き方など、日本語学術ライティング指導の教材は多数出版されているが、その**指導内容は著者の経験蓄積・言語研究による内省で得られたもの**を書いていることがほとんどである。教えられる内容の信頼性に対する学習者の納得感が得られにくい。そこで、日本語学の研究に基づく根拠のある情報を提示し、指導をおこなうことが重要であると考えた。

一方、そのような根拠情報となりうる国立国語研究所開発の現代日本語書き言葉均衡コーパス (BCCWJ) (XML データ形式で総データ量約 1 億語が蓄積されている) が 平成 23 年度末に完成した。研究開始当初の時点では、活用は始まったばかりで、言語学的興味・日本語学的興味からのアプローチに集中していた。貴重な大規模日本語資源を教育に活用するのは重要なことであるが、研究開始当初の時点では、それを大規模に行おうとする事例はなかった。また、現代日本語書き言葉均衡コーパスは、現代日本語の使用実情をなるべく広い範囲から集めたもので、学術的文書のみを蓄積したものではない。一方、学術的ライティングの参考になるような大規模の日本語学術文コーパスは、研究開始時点では存在せず、研究進展に当って我々がデータ構築に携わる必要があると考えた。

本研究の代表者は、これまで大阪大学において、科学研究費その他の研究補助を受け、多数の言語を並列して扱う XML 形式の言語資源を作成研究してきた。本研究において、その研究成果を日本語学習コンテンツ開発の分野に応用することを試みようとの着想に至った。

2. 研究の目的

本研究の目的は、**日本語コーパス（言語資源）から言語特徴抽出を行い、その情報を学習コンテンツとして提示することにより、日本語学術ライティング教育の進歩をはかること**である。

まず、本研究では、日本語学術ライティングの参考教材として活用できそうな、語彙頻度をはじめとする言語特徴情報を現代日本語書き言葉均衡コーパス（BCCWJ）のデータより情報抽出することを目的とする。

続いて、独自開発する学術文コーパスを基礎とすることで、**特定の著者や学会に偏らない言語特徴情報を抽出し**、教材として提供することを目的とする。これにより、より汎用性の高い技能を受講者に身につけさせることができると予想される。

3. 研究の方法

本研究の目的を達成するため、以下の 3 項目の方法による研究推進を計画した。

(1) [**言語情報付与済データの特徴を利用した、BCCWJ データからの語彙情報抽出法研究**]

BCCWJ の XML データのうち、より学術文に近い言語使用状況であると予想される高校教科書データを読み込み、単語頻度情報など、抽出すべきデータを得るためのシステム開発を行う。

(2) [**平文学術文データに対する長単位形態素情報の付与作業法開発**]

平文学術文データから言語特徴情報を抽出するためには、形態素情報など、言語学的情報を付与しなくてはならない。特に専門用語・学術用語を取り出すためには、長単位と呼ばれる形態素単位を採用することが有用である。大量の言語データを効率よく作業できる手法と環境設定法の開発を行う。

(3) [**学術文・技術文データの長単位解析と語彙頻度情報の抽出**] 開発した作業環境を利用し、大阪大学リポジトリで公開されている学術文データ、大阪府立産業技術総合研究所で公開されている研究技術報告文データを対象に、長単位解析と動詞・名詞の語彙頻度情報抽出作業を行う。

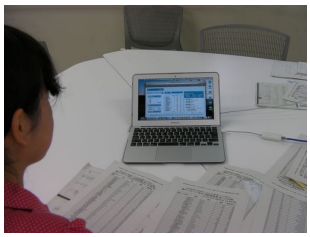
4. 研究成果

(1) [**言語情報付与済データの特徴を利用した、BCCWJ データからの語彙情報抽出法研究**]

主に平成 25 年度の成果として、BCCWJ の高校教科書データの単語頻度情報抽出と、そのライティング授業で提示する参考教材作成を行った。

長単位の形態素情報を XML タグとして付与した形で提供されているデータをデータベースソフトウェア MySQL に格納し、連携して統計処理を行うプログラムを統計処理システム R 上で開発し、処理を行った。得られた高校教科書の一般動詞・普通名詞頻度表リストは、それぞれ頻度上位 300 語分である。これを、利用しやすいよう成形し、大阪大学で開講している（主に学部初年次生を対象とする）日本語学術文ライティング指導セミナー型授業の一

図 1　参考語彙情報を基に学生が学んでいる様子



教材として図 1 のように受講生に提供した。

だだし、この時点で得られた頻度情報は、元データが高校教科書であることに影響されたものとなり、必ずしも大学で学ばせる学術文の規範として直接参考にはなりがたいものであった。このことより、解析の対象は、学術論文・専門技術解説文などより学術文として高度なものを対象にすべきであると判断することになった。

また、本研究を遂行することで得られた知見をもとに、学部初年次生向けの日本語学術文ライティング指導教材である「阪大生のためのアカデミック・ライティング入門」を作成し各年度学部入学者全員に配布している。あわせて、この教材に基づいてライティング教育を行う教員の参考になるよう、教員マニュアルも作成した。本教材は可能な限り高頻度で改訂し、視覚障害を持つ利用者にも便利な大文字白黒反転版も作成した。これらのデータは大阪大学リポジトリ OUKA 上で公開し、広く社会の利用に供している。5. 主な発表論文等の [図書] 欄を参照していただきたい。

(2) [**平文学術文データに対する長単位形態素情報の付与作業法開発**]

2014 年に平文データに対し長単位形態素情報を付与する機能を持つソフトウェア”Comainu”を小澤俊介氏らが開発し、公開された。また、2015 年 3 月には解析精度や活用法の改良がなされた。本研究では、これを利用し、学術文データの長単位形態素情報付与作業の環境構築を行った。

作業環境は Ubuntu Linux 14.04 LTS 上に構築した。Comainu ソフトウェアはシェル上のコマンドとして個別データごとに実行した。得られた長単位情報付与済データから一般動詞あるいは普通名詞とタグ付けされたデータのみを抜き出し、その頻度を計算するシェルスクリプトを作成し、実行した。スクリプト等の詳細は、学会発表データ 8 番の 2016 年 3 月言語処理学会第 22 回全国大会の予稿集に記載している。

(3) [**学術文・技術文データの長単位解析と語彙頻度情報の抽出**]

前記の作業手法により、学術文・技術データの解析と語彙頻度情報の抽出を行った。

学術文として、大阪大学リポジトリ OUKA 上で公開されている博士論文概要を対象とした。分野は、言語学・医学・法学・生物学・教育学のものを 1 件ずつ選び、概要文の箇所のみをテキストデータとして集約した。総字数は約 6,000 字である。一般動詞・普通名詞の頻度表を得た。その動詞頻度表の一部を表 1 に示す。

技術文として、大阪府立産業技術総合研究所の技術報告および技術論文概要の平成 24 年度～平成 27 年度分をテキストデータとして集約した。総字数は約 15,000 字である。一般動詞・普通名詞の頻度表を得た。その動詞頻度表の一部を表 2 に示す。

解析の対象規模は小さく、試行を始めた段階に過ぎないが、長単位による形態素解析をすることで、学術文・技術文の表現特徴をよりよくマイニングすることができるであろうとの予測に至った。

今後の研究進展に向けて

萌芽研究としての本研究の成果を基盤とし、今後さらに大規模・有用な結果が得られるデータ解析作業と手法開発へと進む所存である。

- 解析対象データの大規模化
  各大学が整備を進めているリポジトリに掲載されている論文データを広く対象にすることを予定している。可能であれば、国立情報学研究所の CiNii 論文情報なども対象範囲に含めることを検討している。
- 言語情報抽出用法の改良
  本研究では、コーパス語彙頻度情報をもって特徴抽出とする、簡易な解析手法を採用した。今後より学術文・技術文の言語特徴を効率的に抽出する解析手法を検討し、適用すべく研究進展する。
- 資料インストラクション手法の改善
  大学生および技術者にとって、抽出情報をより

表 1　大阪大学博士論文概要 5 例 (計約 6000 字) から抽出した動詞頻度表 (頻度上位 30 語まで)(堀一成 作成)

長単位動詞	長単位頻度
用いる	12
行う	11
有る	9
する	8
成る	7
存在する	7
割る	7
見る	7
考える	7
示す	6
異なる	6
活性化する	6
燐酸化する	6
着目する	5
調べる	5
発現する	4
減少する	4
議論する	4
示唆する	4
呼ぶ	3
踏まえる	3
構成する	3
描く	3
不活性化する	3
終息する	3
確認する	2
提示する	2
介する	2
応ずる	2
働く	2

表 2　大阪府産技研技術報告概要文集 (計約 15,000 字) から抽出した動詞頻度表 (頻度上位 30 語まで) (堀一成 作成)

長単位動詞	長単位頻度
用いる	35
行う	35
する	21
有る	20
得る	19
成る	16
使用する	12
報告する	11
検討する	10
分かる	10
示す	10
利用する	8
有する	8
含む	8
解説する	7
形成する	7
発生する	7
試みる	6
測定する	6
述べる	6
注目する	6
考える	6
紹介する	6
優れる	5
作製する	5
期待する	5
比べる	5
提案する	4
基づく	4
掛かる	4

参考になる形で提供する手法を継続的に検討していく。

5. 主な発表論文等

〔雑誌論文〕 (計 2 件)

1. 末田真樹子、堀一成、久保山健、坂尻彰宏、職員・教員・TA 協働による学習支援の取り組みー大阪大学附属図書館における「レポートの書き方講座」を中心にー 、大阪大学高等教育研究、査読無、Vol.2、2014、pp.55-60。
   http://hdl.handle.net/11094/28096
2. 堀一成、坂尻彰宏、大阪大学におけるアカデミック・ライティング教育の実践と教材作成、大阪大学高等教育研究、査読無、Vol.3、2015、pp.27-32。
   http://hdl.handle.net/11094/51489

〔学会発表〕 (計 8 件)

1. 堀一成、久保山健、コモンズスペースを利用した教員・図書館職員・TA 協働ライティング指導、大学教育学会 第 35 回年次大会、2013 年 6 月 2 日発表、東北大学 川内キャンパス
2. 堀一成、坂尻彰宏、石島悌、BCCWJ 教科書データより抽出した頻度情報に基づく日本語ライティング指導教材の作成、第 4 回コーパス日本語学ワークショップ、2013 年 9 月 6 日発表、国立国語研究所
3. 堀一成、坂尻彰宏、石島悌、現代日本語書き言葉均衡コーパスより抽出した頻度情報に基づく日本語学術ライティング指導教材の作成、電子情報通信学会 第 3 回テキストマイニングシンポジウム、2013 年 9 月 12 日発表、国立オリンピック記念青少年総合センター
4. 堀一成、坂尻彰宏、全学出動体制を目指したアカデミック・ライティング指導と関連 FD の取り組み、大学教育学会 第 36 回年次大会、2014 年 6 月 1 日発表、名古屋大学 東山キャンパス
5. 堀一成、大学リポジトリデータを活用したアカデミック・ライティング教材の作成、第 16 回図書館総合展、2014 年 11 月 5 日 ~11 月 7 日発表、パシフィコ横浜
6. 堀一成、大阪大学における全学出動体制を目指したアカデミック・ライティング指導の取り組み、大学教育改革フォーラム in 東海 2015、2015 年 3 月 7 日発表、名古屋大学 ES 総合館
7. 堀一成、山内保典、家島明彦、浦田悠、大学院生向け Transferable Skills Workshop の可能性: 大阪大学の挑戦、大学教育学会 第 37 回年次大会、2015 年 6 月 7 日発表、長崎大学 文教キャンパス
8. 堀一成、坂尻彰宏、石島悌、ライティング教材作成を目指した日本語学術文長単位解析の試行、言語処理学会 第 22 回全国大会、2016 年 3 月 9 日発表、東北大学 川内キャンパス

〔図書〕 (計 7 件)

1. 堀一成、坂尻彰宏、大阪大学のためのアカデミック・ライティング入門、2014、総ページ数 32。
   http://hdl.handle.net/11094/27153
2. 堀一成、坂尻彰宏、「大阪大学のためのアカデミック・ライティング入門」ライティング指導教員マニュアル、2014、総ページ数 14。
   http://hdl.handle.net/11094/27594
3. 堀一成、坂尻彰宏、大阪大学のためのアカデミック・ライティング入門 第 2 版、2015、総ページ数 32。
   http://hdl.handle.net/11094/51131
4. 堀一成、坂尻彰宏、「大阪大学のためのアカデミック・ライティング入門」ライティング指導教員マニュアル Ver.2、2015、総ページ数 14。
   http://hdl.handle.net/11094/51132
5. 堀一成、坂尻彰宏、大阪大学のためのアカデミック・ライティング入門 第 3 版、2016、総ページ数 32。
   http://hdl.handle.net/11094/54512
6. 堀一成、坂尻彰宏、大阪大学のためのアカデミック・ライティング入門 第 3 版 白黒反転大文字版、2016、総ページ数 95。
   http://hdl.handle.net/11094/54607

7. 堀一成、坂尻彰宏、「大阪大学のためのアカデミック・ライティング入門」ライティング指導教員マニュアル Ver.3、2016、総ページ数 19。
http://hdl.handle.net/11094/54513

〔その他〕
ホームページ
大阪大学　学術情報庫 OUKA 阪大生のためのアカデミック・ライティング情報公開 Web ページ
http://ir.library.osaka-u.ac.jp/web/HAW/index.html

6. 研究組織

(1) 研究代表者
堀　一成 (Hori Kazunari)
大阪大学・全学教育推進機構・准教授
研究者番号：80270346

(2) 研究分担者
坂尻　彰宏 (Sakajiri Akihiro)
大阪大学・全学教育推進機構・准教授
研究者番号：30512933

石島　悌 (Ishijima Dai)
大阪府立産業技術総合研究所・
製品信頼性科・主任研究員
研究者番号：80359398