様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月　 ９ 日現在

機関番号：１４４０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ０１０１６

研究課題名（和文）学術コーパスから抽出した情報に基づく科学技術ライティング指導教材作成法の研究

研究課題名（英文）Research on providing the learning contents for science academic writing based on the mined data from research paper 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）　　3,500,000円

研究成果の概要（和文）：　科学技術系の大阪大学公開日本語博士論文データや公立技術研究所の技術解説文データから言語特徴抽出を行った。また、日本語科学技術アカデミック・ライティング教育のための動画教材開発や対面講習の実践を行った。
　言語特徴抽出の対象とした博士論文データの分量は約450万字、技術解説文データの分量は約35万字である。それぞれのデータに対する長単位形態素情報の付与作業法開発と語彙頻度情報の抽出作業を行った。

研究成果の学術的意義や社会的意義
　本研究の成果により、特に科学技術分野のアカデミック・ライティングで用いることを推奨する表現を科学的に検証可能な形で提供できた。これは日本語科学技術学習教材開発研究においては従来になかった手法であり、データ抽出法手順を開発することで日本語コーパス研究の進展にも貢献したといえる。また、信頼のおける情報を学習教材として提供できるようになることから、大学科学リテラシー教育の新展開を可能とする成果であるといえる。

研究成果の概要（英文）：We extracted the language features from the doctoral dissertations in science and technology published by Osaka University and the technical reports from public research institutes of industrial science and technology. We also developed video teaching materials and practiced face-to-face lessons to teach Japanese academic writing in the field of science and technology.
The doctoral dissertations data used for language feature extraction included approximately 4.5 million Japanese characters, and the technical reports had approximately 350,000 Japanese characters. We developed a method of assigning long-unit morpheme information and extracted vocabulary frequency information.

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　科学教育　教材開発　学術コーパス　技術文書　自然言語処理　長単位形態素解析

様　式 C–１９、F–１９–１、Z–１９（共通）

1. 研究開始当初の背景

日本語科学技術系学術文章データを情報基盤とし、言語学・日本語学の研究に基づく根拠のある情報を提示し、アカデミック・ライティング指導をおこなうことが大学や研究機関において求められている状況であった。従来、著名な木下是雄著『理科系の作文技術』[1] をはじめとして、日本語科学技術系アカデミック・ライティング指導の教材が数多く出版されている。しかし、その指導内容は著者の経験蓄積によるもの (言語学研究分野ではこれを著者の内省と呼ぶ)、あるいは欧米の研究者向けライティング指導書の内容を日本語化したものを紹介していることがほとんどである。このような状況では、日本語科学技術系アカデミック・ライティング指導の内容に対する信頼性に関して、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、（特に科学技術系の）アカデミック・ライティング指導をおこなうことが重要であると考えた。

一方、そのような日本語学の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。順次活用が進んでいるが、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、学術文書の割合は非常に少ない。一方、アカデミック・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 形式のコーパスを作成研究してきた。また、コーパスデータを統計処理し、アカデミック・ライティング指導教材の素材となる情報を抽出する試みを重ねてきた。本研究において、その研究成果を日本語科学技術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、**大規模なコーパス（言語資源）から言語特徴抽出を行い、その情報を学習教材として提示し、対応するカリキュラムを開発することにより、日本語科学技術アカデミック・ライティング教育の進歩をはかること**である。

本研究では、大学・研究所で蓄積した日本語科学技術文章を基に学術文章コーパスを構築する。そこから得られた特徴情報に基づく教材・カリキュラムを開発することで、**特定の著者や学会に偏らない、より汎用性の高いアカデミック・ライティング技能を受講者に身につけさせること**を目指している。

3. 研究の方法

本研究の目的を達成するため、以下の 2 項目の研究を計画した。

(1)[**学術文章データのコーパス化作業と集積**] 大学・研究所で蓄積した科学技術学術文章データを選定し、学術文コーパス化の作業を行う。情報抽出と教材化が容易になるよう、データベース等に蓄積する。特に次の２段階の作業が重要である。

(A)[**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**]

Web で公開されている学術文データは、（言語学的特徴情報が付与されていない）平文テキストデータであることがほとんどである。平文テキストデータから言語特徴情報を抽出するためには、形態素解析など、言語学的情報を抽出する処理をする必要がある。特に（本研究で重要となる）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に基づいた形態素解析をすることが重要である。しかし、多くの形態素解析ソフトウェアは、短単位と呼ばれるできるだけ短い要素の形態素を抽出することを主機能としており、長単位形態素データを獲得するためには、ソフトウェアの選定や作業手順の工夫が必要である。本研究以前の研究においても、手法の開発はおこなってきたが、改めて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。

(B) [**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、大阪産業技術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。

(2)[**構築したデータを科学技術ライティング教育へ応用する実践研究**] コーパスからの抽出情報を基に、特に科学技術文章に特有の語彙・連語・文体などの言語的特徴情報を整理し、動画を含むアカデミック・ライティング指導の教材の素材とする。指導者のためのマニュアルやシラバス案なども併せて作成し、科学教育カリキュラムを構築する。大学におけるアカデミック・ライティング授業や研究所内研修で活用することで、成果を確認する。

4. 研究成果

(1)[**学術文章データのコーパス化作業と集積**]

(A) [**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**]

長単位形態素解析ソフトウェアの選定は、研究期間中に新しく発表されたソフトウェアを含め、数種の特質を検討した。その結果、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として、前研究における試行でも利用した、小澤俊介氏らが開発し、公開されている”Comainu-0.72”[2] を採用することとした。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 18.04 LTS、および macOS 10.13.6 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。まず、Web 上からダウンロードした論文等データはすべて PDF 形式であるので、PDF データから書かれている文章のテキストデータを抜き出す処理を行なった。この処理は、Java 言語の Apache Tika ライブラリを利用することで実現した。得られたテキストデータには、不要な空白・改行・記号などがあるため、これを取り除き、長単位解析がスムーズに行えるようにするための前処理ソフトウェアを AWK 言語で開発した。前処理済データを Comainu で長単位形態素情報付与済データに分解した。

(B) [**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**]

(A) の作業で得られた長単位形態素情報付与済データから、品詞情報が「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 Python プログラムを開発し、実行することにより頻度データを得た。

一連の作業手順に従って、学術文および研究技術報告文データの解析と語彙頻度情報の抽出を行った。

科学技術系学術文として、大阪大学リポジトリ OUKA 上で公開されている日本語で本文が書かれた博士論文を対象とした。分野は、理学・工学・医学・薬学などの分野で、2010 年度以降に学位申請され、公開された論文を、ランダムに 107 件選びだした。PDF からテキストデータ化したデータ量は、全角文字数で約 450 万字となった。このデータを処理し、一般動詞・普通名詞の頻度表を得た。より専門的な語彙を抽出するため、国立国語研究所の国語研教育基本語彙 [3] のうち、特に基本的とされる 2000 語に含まれる動詞を除く処理を行った。得られた動詞頻度表の頻度上位データの一部（30 語）を表 1 に示す。

研究技術報告文として、大阪産業技術研究所が Web ページで公開している、テクニカルシート 463 文書、技術資料 23 文書を解析対象とした。PDF からテキストデータ化したデータ量は、全角文字数で約 35 万字となった。科学技術系学術文と同様の処理行い、（基本語彙を除いた）一般動詞・普通名詞の頻度表を得た。得られた動詞頻度表の頻度上位データの一部（30 語）を表 2 に示す。

本研究の代表者・分担者の個人的な印象であるが、科学技術系学術文の表現特徴を一定程度抽出できたと判断している。

表 1　大阪大学科学技術分野博士論文本文データ 107 例 (計約 450 万字) から抽出した動詞頻度表 (基本語彙を除く頻度上位 30 語まで)(堀一成 作成)

長単位動詞	長単位頻度
得る	1007
示せる	468
変化する	354
表わす	316
置ける	303
使用する	275
発生する	272
異なる	247
報告する	228
形成する	220
生じる	206
提案する	206
実施する	205
測定する	194
存在する	177
利用する	175
増加する	174
比較する	170
内包する	169
考慮する	168
減少する	162
有する	160
生ずる	153
構成する	145
開発する	143
検討する	140
作製する	139
確認する	137
評価する	135
実現する	128

表 2　大阪産業技術研究所技術解説関連文 486 例 (計約 35 万字) から抽出した動詞頻度表 (基本語彙を除く頻度上位 30 語まで) (堀一成 作成)

長単位動詞	長単位頻度
得る	208
測定する	172
異なる	126
利用する	116
発生する	107
紹介する	102
使用する	90
優れる	73
変化する	72
存在する	61
応ずる	60
作製する	60
有する	57
増加する	57
生じる	57
表わす	52
形成する	51
生ずる	51
及ぼす	47
対応する	47
分析する	46
伴う	43
導入する	43
評価する	43
溶解する	40
低下する	40
算出する	38
照射する	37
測定出来る	37
腐食する	36

(2)[**構築したデータを科学技術ライティング教育へ応用する実践研究**]

これまでの研究成果を科学技術ライティング教育において活用するため、教育実践を試みた。

まず、高校生を対象とした日本語科学アカデミック・ライティング講習を実践した。大阪大学では、大学で行われている科学技術研究を、高校生に体験してもらう SEEDS プログラムと称する学習プログラムを、2015 年度から継続して提供している。本研究の代表者（堀）は、「アカデミック・ライティング ＝ 科学技術編 ＝」と題する講習会を SEEDS 参加の高校生を対象として、2016 年度以降毎年 1 回行っている。この講習会の様子を図 1 に示す。ただし、本研究で得られた学術語彙頻度表については、この講習会の教材として提示することができなかった。

図 1　大阪大学 SEEDS 参加の高校生に対し、研究代表者 (堀) が科学ライティングを講習している様子

また、学部初年次生向けの日本語アカデミック・ライティング指導を内容とする動画教材を作成した。作成済みの教材冊子「阪大生のためのアカデミック・ライティング入門」に基づいてアカデミック・ライティングの基本事項を解説するものである。本報告の時点で作成できた教材の分量は、約 30 分の動画が 6 本であり、講師役を本研究の代表者（堀）と分担者（坂尻）が務めている。これらの動画教材データは Youtube 上の「阪大ウェルカムチャンネル」の一部として利用に供している。ただし、動画での説明内容は一般の日本語アカデミック・ライティング全般にわたるもので、科学技術分野に特化したものとはなっていない。

当初予定していた、公立研究所の研究員を対象とした研究技術報告文作成のための研修は、研究期間中に実施できなかった。

図 2　Youtube で配信した大阪大学初年次生向け学習動画教材において、研究分担者 (坂尻) がアカデミック・ライティングを講習している様子

今後の研究進展に向けて

本研究の成果を基盤とし、さらに対象者が広く学習効果の高い科学技術系アカデミック・ライティング教育に関する研究へと進む所存である。

- 解析対象データの大規模化
  今回、大阪大学公開の博士論文のみを解析対象としたが、他大学が整備を進めてリポジトリ公開している論文データなども対象にすることで、データのさらなる大規模化をはかることができる。特に、科学技術振興機構の J-STAGE で公開されている論文情報を対象範囲に含めることを検討している。
- 言語特徴情報抽出法のさらなる改良
  本研究では、形態素別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
  学習者にとって（特に学術文の作成を初めて体験する高校生にとって)、研究成果情報をより参考になる形で提供する教材のありかたや説明手法を高大接続の視点に立って、継続的に開発していく。

引用文献

[1] 木下是雄『理科系の作文技術』中央公論社 (1981).
[2] 小澤俊介、内本清貴、伝康晴「BCCWJ に基づく長単位解析ツール Comainu」言語処理学会 第 20 回年次大会 発表論文集 (2014)、pp.582-585.
[3] 国立国語研究所『教育基本語彙の基本的研究』国立国語研究所報告 117(2001).

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　4件）

１．著者名 村岡貴子・堀一成・坂尻彰宏	４．巻 22
２．論文標題 大阪大学における日本語ライティング教育の実践	５．発行年 2018年
３．雑誌名 多文化社会と留学生交流	６．最初と最後の頁 23-32
掲載論文のDOI（デジタルオブジェクト識別子） info:doi/10.18910/67904	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 根岸千悠、坂尻彰宏、堀一成、山口和也	４．巻 5
２．論文標題 初年次教育科目としての理系学生対象アカデミック・ライティングの授業デザイン	５．発行年 2017年
３．雑誌名 大阪大学高等教育研究	６．最初と最後の頁 63-69
掲載論文のDOI（デジタルオブジェクト識別子） info:doi/10.18910/60492	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 吉本真代、和嶋雄一郎、坂尻彰宏、堀一成	４．巻 8
２．論文標題 大学入学者の高校での『書く』経験は変化しているのか	５．発行年 2020年
３．雑誌名 大阪大学高等教育研究	６．最初と最後の頁 13-19
掲載論文のDOI（デジタルオブジェクト識別子） info:doi/10.18910/75496	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 堀一成、坂尻彰宏、進藤修一、柿澤寿信、金泓樘、田中誠樹、竹林祥子、大泉幸寛、宮崎雄史郎	４．巻 8
２．論文標題 高大連携により取り組む高校生に対するアカデミック・ライティング教育の実践	５．発行年 2020年
３．雑誌名 大阪大学高等教育研究	６．最初と最後の頁 51-60
掲載論文のDOI（デジタルオブジェクト識別子） info:doi/10.18910/75500	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　0件）

１．発表者名
堀一成・坂尻彰宏

２．発表標題
大阪大学2016年度学部新入生アカデミック・スキル調査

３．学会等名
第2回 大阪大学　豊中地区研究交流会

４．発表年
2018年

１．発表者名
堀一成・坂尻彰宏・齋藤貴浩

２．発表標題
大阪大学の学部新入生に対する日本語ライティングスキル大規模調査

３．学会等名
第24回　大学教育研究フォーラム

４．発表年
2018年

１．発表者名
堀一成,坂尻彰宏,太田ユカ,奥村典夫,中藤強,松井佳代美

２．発表標題
高大連携と大学院生キャリア開発を重視した高校生に対する日本語アカデミック・ライティング指導

３．学会等名
第23回 大学教育研究フォーラム

４．発表年
2017年

１．発表者名
堀一成

２．発表標題
日本語ライティング支援者の養成 ―大阪大学における大学院生科目での実践―

３．学会等名
2019年度日本語教育学会春季大会

４．発表年
2019年

１．発表者名
堀一成、吉本真代、和嶋雄一郎 、坂尻彰宏

２．発表標題
高校の探究学習が大学入学者のライティングスキルに与える影響　―大阪大学入学時アンケートより―

３．学会等名
第26回大学教育研究フォーラム

４．発表年
2020年

〔図書〕　計2件

１．著者名 堀一成	４．発行年 2018年
２．出版社 くろしお出版	５．総ページ数 223
３．書名 大学と社会をつなぐライティング教育	

１．著者名 堀一成、坂尻彰宏	４．発行年 2019年
２．出版社 大阪大学　全学教育推進機構	５．総ページ数 32
３．書名 阪大生のためのアカデミック・ライティング入門　第４版	

〔産業財産権〕

〔その他〕

大阪大学　全学教育推進機構　学生支援「アカデミック・ライティング」ページ
https://www.celas.osaka-u.ac.jp/education/support/academic-writing/
阪大ウェルカムチャンネル
https://www.youtube.com/channel/UCa3nLV_BiehQlbPAnpSm2Kw

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	坂尻　彰宏 (SAKAJIRI Akihiro) (30512933)	大阪大学・全学教育推進機構・准教授 (14401)	
研究分担者	石島　悌 (Ishijima Dai) (80359398)	地方独立行政法人大阪産業技術研究所・製品信頼性研究部・主任研究員 (84415)	削除：2018年12月26日