科学技術系ライティング教材作成のための Comainu を利用した日本語学術文技術文長単位解析

堀 一成 *, 坂尻 彰宏 (大阪大学 全学教育推進機構)

Long-unit-word morphological analysis on Japanese academic and technical research paper corpora using Comainu for providing the learning contents for science academic writing

HORI Kazunari, SAKAJIRI Akihiro (Osaka University)

要旨

大学学部初年次生向け科学技術系日本語アカデミック・ライティング指導教材を作成する際の基礎データとするため、学術文・技術文の長単位による形態素解析を行い、用いられている（基本語彙を除く）一般動詞の頻度情報を得た。長単位形態素解析に用いたソフトウェアは、小澤俊介氏らの開発した Comainu-0.72 を採用した。学術文の代表として、大阪大学に提出された理学・工学・医学・薬学などの日本語博士学位論文の本文（107 件でデータ量は、全角文字数で約 450 万字）を、技術文の代表として、大阪産業技術研究所が公開している技術報告文 (486 文書、データ量全角文字数で約 35 万字) を、解析の対象として選定した。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も行った。本報告では、研究の背景、ソフトウェア実行手順、得られた成果に対する考察などを紹介する。

1. 研究の背景

大学学部初年次生を対象とした日本語アカデミック・ライティング指導教材は多数発行されている。報告者らも独自の教材を作成し、大阪大学の学部初年次生が自由に利用できるようデータ公開している (堀・坂尻 (2020), 堀・坂尻 (2015))。しかし、その指導書などでは、文を書く際に使用する用語や言い回しの事例紹介がなされる例が多いが、その用例・文例を採用した根拠が明示されていることはまれである。一般に日本語教材の説明は著者の内省によっており、その根拠となる情報を示されることは少ない。（二通ほか著「レポート・論文表現ハンドブック」:東京大学出版会 (二通ほか (2009)) は少ない例外であるが、これも心理・経済分野の特定著者の論文が論拠である。）

2016 年に、我々は、今回の報告の基盤となる、科学技術分野の学術文・技術文の解析試行の成果を報告した (堀ほか (2016))。この研究の背景には、特定の著者や学会に偏らないデータが

* hori[アットマーク]celas.osaka-u.ac.jp

得られ、その 成果をライティング指導に活用することで、特に科学技術分野のより広範囲に応用できるライティング技能を受講者に身につけさせることができるとの着想がある。

これは、2014 年に平文テキストデータを入力とし、専門用語など学術文に特徴的な語彙を抽出できる、長単位解析可能な形態素解析ソフトウェア Comainu と、関連した形態素辞書である Unidic2 が公開され、長単位基準の多様な言語資源に対する特徴語抽出が可能になったことが研究進展の大きな要因であった。

2016 年の報告は、試行であり、解析対象データ規模が学術文については約 6,000 字、技術文については、約 15,000 字と小規模なものであった。そのため、得られた結果が言語的特徴を十分に抽出できているとは言えなかった。より有効な教材データを得るためには、大規模なデータに基づく解析が必要であった。

1.1 長単位解析ソフトウェア Comainu

小澤俊介氏らが開発している形態素解析ソフトウェア Comainu(小澤ほか (2014)) は、その解析結果出力単位に長単位を選ぶことができることが、大きな特徴である。冨士池ら (冨士池ほか (2008)) によると、長単位とは、文節の内部を自立語部分と付属語部分に分解することで認定される区切りである。長単位は資料の特徴語を取り出せることが利点であるとしている。対して、短単位は基準がわかりやすく、ゆれが少ないが、合成語を構成要素に分割してしまう問題点があるとしている。

2. 頻度リストの作成方法

以下に長単位の動詞の頻度リストを作成した手順を説明する。作業は CPU Intel Xeon 6-Core 3.33GHz, メモリ 6GB 搭載の Mac Pro (2010) 上で行った。OS は Mac OS 10.13.6(High Sierra) である。

1. **解析対象平文データの準備**

   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。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。PDF から UTF-8 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。これは 2016 年報告の試行時の 1000 倍近い文字数となっている。

   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。学術文と同様の処理を経て、PDF から UTF-8 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。こちらは、2016 年報告の試行時の 20 倍強の文字数となっている。

2. **Comainu を用いた解析作業**

上記作業により得た UTF-8 テキストデータを Comainu Ver.0.72 で、平文から長単位の解析結果が得られるようオプション設定して処理する。

3. **Python による頻度情報抽出作業**

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」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動詞頻度データを得た。

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 (国立国語研究所 (2001)) 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。最後に、作業者（堀）が目視判断により、大学学部初年次生に提供することが適当と考える語彙を残す作業を行った。これらの最終成形作業は Excel2019 上でおこなった。

このようにして得られた学術文と技術文に含まれる一般動詞のうち、基本的な語彙を除いたもの頻度を、表 1 と表 2 に示す。いずれも頻度上位 30 位までの動詞データである。

3. 得られた頻度リストに対する考察

得られた頻度リストのうち、一般動詞データに対して考察する。

印象的説明になるが、表 1 と表 2 に示した 2 例ともに、長単位解析により、「抽象名詞＋する」動詞を中心とした、学術的な硬い表現がマイニングできているといえる。このことから、普段このような表現を書きなれない学部初年次生が日本語アカデミック・ライティング時に使用するよう推奨する語彙を検討する場合、選定の基準に関する有用な示唆を長単位解析結果が与え得る。

また、これも印象的記述にすぎないが、博士学位論文データと大阪産技研データの長単位リストを比較すると、大阪産技研データの方に、より産業技術説明分野に特化した、表現が多いようである。

4. 今後の展開

本報告は、2016 年に報告した手法を拡大適用し、学術文・技術文の長単位解析データを有効活用するための一例を紹介したものとなっている。今回の成果をふまえ、さらに大規模・有用な結果がえられる手法開発へと進みたいと考えている。

◎ 解析対象コーパスデータの大規模化

本報告の学術文データは、大阪大学に提出された博士論文データに限ったものであった。解析対象コーパスデータをさらに大規模なものとするため、大阪大学以外の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。さらに、科学技術振興機構の J-Stage に収録されている論文情報なども対象にすべきと考えている。技術文解析対象範囲についても、大阪産業技術研究所以外のデータも対象とするよう努めていく。

◎ 特徴的な語・表現の抽出方法の改良

今回、特徴語の抽出方法は、長単位解析して頻度情報を得るだけという、簡易な手法であった。今後適切な、学術文・技術文と一般分のデータ集団の言語特徴差異抽出手法を検討し、よ

り良い抽出結果を得たいと考えている。単なる語彙情報のみに注目するのではなく、連語 (コロケーション) の情報の提供がより有用であろうと予想している。

◎ 高校生も対象に入れたインストラクション手法の改善

日本語アカデミック・ライティングの学習者にとってより参考になるよう、研究成果情報を提供する教材のありかたや説明手法を継続的に開発していく。報告者らが 2020 年度現在において注力している高大接続活動において、(特に学術文の作成を初めて体験する高校生にとって) 学習者に、今回作成した動詞頻度データのみを渡すだけでは、有効活用は期待できない。頻度データに加え、関連の言語学的情報や関連 Web ツールを使用して、より良い表現を選定する具体的な方法を、文章作成指導手順に組み込み提示したいと考えている。

既存の教材も、研究成果に基づき、改善する作業を進行させる予定である。

5. おわりに

以上のように、大学学部初年次生を主な対象として、科学技術分野の日本語アカデミック・ライティング指導教材を作成する際の基礎データとするため、学術文・技術文の長単位による形態素解析を行い、用いられている一般動詞のうち基本語彙を除いたものの頻度情報を得た。学術文の代表として、大阪大学に提出された科学技術分野の日本語博士学位論文の本文データを、技術文の代表として、大阪産業技術研究所の研究技術報告文を、解析の対象として選定した。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ことが確認できた。

謝 辞

本研究は、学術研究助成基金助成金基盤研究 (C) 課題番号: 16K01016「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」(研究代表者：堀一成) および、基盤研究 (C) 課題番号: 20K03251「『ダメな科学ライティング』をさせないための高大接続による探究学習教育法の研究」(研究代表者：堀一成) による補助を受け推進しているものである。

本研究は、小澤俊介氏を代表とする Comainu 開発グループの成果に依存したものである。有用なソフトウェアの開発と公開に対して深く謝意を表したい。

文 献

堀一成・坂尻彰宏 (2020).『阪大生のための アカデミック・ライティング入門第 4 版』 大阪大学 全学教育推進機構 http://hdl.handle.net/11094/71454 から自由に PDF ファイルをダウンロードできる

堀一成・坂尻彰宏 (2015).「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 Vol.3, pp. 27–32.

二通信子・大島弥生・佐藤勢紀子・因京子・山本富美子 (2009).『留学生と日本人学生のためのレポート・論文表現ハンドブック』 東京大学出版会.

堀一成・坂尻彰宏・石島悌 (2016).「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 回年次大会発表論文集, pp. 685–688.

小澤俊介・内元清貴・伝康晴 (2014).「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, pp. 582–351.

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 (2008).「「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 回年次大会発表論文集, pp. 931–934.

国立国語研究所 (2001).『国立国語研究所報告 117 教育基本語彙の基本的研究』 国立国語研究所.

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)
科学研究費基盤研究 (C) 課題番号:16K01016 研究成果報告書より転載

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研究技術報告文 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)
科学研究費基盤研究 (C) 課題番号:16K01016 研究成果報告書より転載

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