科学技術系ライティング教育改善を目的としたシラバス分析のためのspaCY-GSDLUWを利用した日本語長単位解析

科学技術系ライティング教育改善を目的とした シラバス分析のための spaCy-GSDLUW を利用した日本語長単位解析

堀 一成, (大阪大学)*

Long-unit-word morphological analysis on Japanese university syllabus data using spaCy-GSDLUW for providing the learning contents for science academic writing

HORI Kazunari, (Osaka University)

要旨

日本語アカデミック・ライティング科目のシラバスデータを新しい形態素解析ソフトウェア sapCy_GSDLUW で長単位形態素解析することの有用性が明らかになった。発表者らは、大学学部初年次生や高校生を対象とする科学技術系日本語アカデミック・ライティング指導を改善するため、日本語アカデミック・ライティング科目のシラバスデータを収集し、内容の分析を試みている。その際、シラバスデータの言語特徴を良く抽出するため、長単位形態素解析を行うことが有効ではないかと考えている。2022 年に公表された長単位解析ソフトウェア spaCy-GSDLUW は、導入作業も容易で、分析の精度もこれまでのシステムより高いものである。発表者が開講する少数の科目シラバスのみを対象にした試行結果であるが、教育関係情報の言語処理を進めるにあたり、まず長単位形態素解析ことが有用であろうとの目処がついた。試行の方法や得られた結果に対する考察を紹介する。

1. 発表の背景

発表者らは、大学学部初年次生や探究学習に取り組む高校生を対象に、日本語アカデミック・ライティング指導を継続している。また、その指導内容が順次改善されるよう、様々な試みを行ってきた。

大学学部初年次生を対象とした独自の日本語アカデミック・ライティング指導教材を作成し、（所属する大阪大学だけでなく）広く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))。また科学実験レポート作成に特化した理系学部初年次対象のガイドブック (堀ほか (2022)) なども上梓した。

このような指導を行うにあたり、指導内容の根拠となる情報がコーパス分析の結果から得られないかと考え、科学技術分野の学術文・技術文の解析を行った (堀ほか (2016))。その際、学

* hori.kazunari.celas[アットマーク]osaka-u.ac.jp

術文に特有な言語表現を抽出するためには、長単位に基づく形態素解析を行うべきだと判断し、実行し一定の成果を得た。

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。

一方、2016 年より、発表者の所属する研究グループは、大学・大学院に在籍する日本語学習者による読解・ライティングの学習方法や文章観とその背景を、探索的調査から明らかにすることに取り組んでいる。その取り組みの一つとして（日本語非母語話者に対するものを含む）多くの大学で開講されている日本語アカデミック・ライティング指導科目のシラバスデータを収集分析し、ライティング教育の現況を把握する作業を進めている。グループの先行して研究を進めていた研究者は、導入が容易で操作も簡便な KH-Coder を主な分析ツールとして利用していた。KH-Coder は、形態素解析を MeCab か Chasen で実行する設定であり、その後の様々な解析は短単位形態素に基づくものになっている。発表者としては、シラバスに含有されているライティング教育の特徴をよく表す語彙は、長単位形態素に基づくものであろうと予想した。しかし、長単位形態素解析のソフトウェアは普段 Linux など UNIX 系システムを使い慣れないものには扱いが難しいところがあり、積極的な導入進言に逡巡する状況であった。

2022 年に、後述する新しい新しい長単位形態素解析ソフトウェア spaCy_GSDLUW が発表され、これを活用することでシラバスデータ分析の高度化が図れるのではないかとの発想に至った。

本発表は、取り組んでいるシラバスデータ分析といった教育関係情報の言語分析をする際（特に科学技術系ライティング教育の改善に資するような情報が得たい）に、長単位形態素解析システム spaCy_GSDLUW を利用することが有用であるか、目処をつけるための試行の結果を報告するものである。

1.1 長単位解析ソフトウェア spaCy-GSDLUW

国立国語研究所と株式会社リクルート Megagon Labs が共同開発し、2022 年に一般に公開された新しい長単位形態素解析ソフトウェア spaCy_GSDLUW(松田ほか (2022)) は、Universal Dependencies に基づく依存構造解析モデルに基づくものであり、従来手法を上回る長単位品詞推定精度を誇るものである。

また、Python (バージョン 3.10 が推奨されている) と pip システムが動作できるようであれば、pip コマンドを利用し、一行コマンドを入力するだけで GitHub のリポジトリからダウンロード・インストールが完了するため、導入が比較的容易であるところも特徴である。

（注意：本稿作成時点の 2023 年 8 月において、spaCy_GSDLUW は Python3.11 以降のバージョンでは正常動作しない）

2. 頻度リストの作成方法

以下にシラバス分析長単位形態素の頻度リストを作成した手順を説明する。作業は Mac Mini (2020)(Intel Core i7-8700B 3.2GHz) 上で行った。OS は macOS Ventura 13.5 である。

1. **試行対象シラバス平文データの準備**

   試行分析対象として、発表者（堀）が大阪大学で 2023 年度に開講しているセミナー型アカデミック・ライティング指導科目２科目の日本語シラバスを用いた。一つは学部初年次生対象の「学術的文章の作法」であり、もう一つは大学院生対象の「学術的文章の作法とその指導」である。大阪大学の教学システム KOAN 上で提供している両科目のシラバスデータをダウンロードし、結合して改めて UTF-8 テキストデータとした。このデータを Microsoft 365 Word の文字カウント機能でカウントした、スペースを含めない文字数は、5,254 字である。

   本来、形態素解析がより有効に行えるよう、記号・数字・本文内容に関係ない文字列などを取り除く前処理をすべきであるが、今回は作業省力化のため、このような処理を行っていない。

2. **spaCy_GSDLUW と MeCab を用いた形態素解析作業**

   上記作業により得た UTF-8 テキストデータを Python 3.10.10 上で spaCy_GSDLUW を利用し長単位形態素リストを得る処理をした。入力したテキストファイルを、１行ずつ sapCY_GSDLUW ライブラリに渡し、長単位形態素解析したリスト情報を作成し、標準出力に出力する Python プログラムを作成し、実行した。同様の処理を MeCab(Unidic 辞書) を用いて短単位形態素リスト情報を得る Python プログラムを作成し、実行した。

3. **Python による頻度情報抽出作業と Excel によるデータ整理**

   得られた spaCy_GSDLUW の出力である長単位形態素情報付与済データから、各行の形態素原形情報が出力される第 3 列のデータのみを抜き出し、その頻度を計算する Python プログラムを作成し、実行することにより（試行対象であるシラバスデータの）長単位形態素頻度データを得た。

   まったく同様の操作を MeCab の処理対象についても行い、短単位形態素頻度データを得た。

   最後に、作業者（堀）が目視判断により、出力形態素を一般名詞と一般動詞に分け、さらに短単位では分解され単独形態素として抽出できないであろう「漢語＋する」動詞を分離し、別表データとする作業を行った。これらのデータ整理作業は Microsoft 365 Excel 上で手作業で行った。

このようにして得られたシラバス試行分析形態素頻度を、表 1 から表 5 に示す。

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 に示した動詞の長単位・短単位を見比べると、長単位解析

により、「抽象名詞（多くの場合漢語）＋する」動詞を含む、より言語教育に関連した？ 表現が抽出できているといえる。特に短単位で「為る」の頻度が非常に高いことは、上記「〜する」動詞が分解されてしまって、一単語として抽出できていないことを示唆している。特に短単位では頻度リストに出てこず、長単位では出てきた「漢語＋する」動詞を別表である表 5 に提示した。

同様に、表 3 と表 4 に示した名詞の長単位・短単位を見比べると、長単位解析により、「担当教員」「教室対面」「５パラグラフ論証文」など、よりシラバスらしい？ アカデミック・ライティング教育らしい？ 表現が抽出できているといえる。短単位名詞頻度表の最後に「オン」とあるが、長単位名詞頻度表と比べると、これは「オンデマンド」が分解されてしまったものであることがわかる。

このことから、シラバスのような教育関連情報を言語分析するにあたり、その後どのような統計処理をする場合であっても、第一処理段階の形態素解析の処理を長単位に基づいたものとすることは、（短単位処理したデータを基に処理進行することと比べ）おそらく有用な選択であろうということを示唆するものと、発表者は考える。

4. 今後の展開

本発表は、シラバスデータの言語分析の第一処理に、長単位に基づく形態素解析が（短単位に基づくものに比べ）有用であるか、目処をつけるための試行報告であった。

◎ 解析対象日本語シラバスデータの継続収集とコーパス化

本報告の形態素解析対象データは、大阪大学に所属する発表者（堀）個人の開講科目シラバスデータに限ったものであった。本来の目的である、多様なアカデミック・ライティング教育の現状をシラバス分析により把握し、改善の参考データとするためには、できる限り広範なシラバスデータを収集し、分析を進める必要がある。

◎ シラバスデータからの有用な教育改善情報抽出方法の研究

今回は、少数のシラバスデータを長単位形態素解析して頻度情報を得るという、簡易な作業を行っただけである。今後アカデミック・ライティング教育の改善に資する情報を得るためには（長単位形態素解析したデータを出発点とし）どのような統計処理・ＡＩ利用分析などをすればよいか、有効な分析法の研究を進める必要がある。

2023 年 8 月の発表時点において、発表者らが所属する研究チームにおいて、このようなデータ収集法、有効な分析法について議論を継続している。

5. おわりに

以上のように、日本語アカデミック・ライティング科目のシラバスデータを spaCy_GSDLUW で長単位形態素解析することにより、有用な情報を得ることができる目処がついたといえる。2022 年に公表された新しい長単位解析システム spaCy-GSDLUW は、導入作業も容易で、分析の精度も従来の長単位解析システムより高いものである。

本発表の事例を参考にされ、多くの研究で長単位形態素解析が利用されることを望むものである。

謝　辞

本発表の研究は、科学研究費基盤研究 (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）、および基盤研究 (B) 課題番号: 19H01269「日本語読解・ライティングの方法に影響する母語・母文化の教育的背景要因に関する研究」（研究代表者：村岡貴子）による補助を受け推進しているものである。

本発表は、松田 寛氏を代表とする株式会社リクルート Megagon Labs の研究グループと国立国語研究所言語資源開発センターの共同成果である spaCy_GSDLUW 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。

また、日本語非母語話者も対象に含む日本語アカデミック・ライティング教育について、シラバス分析の方針などについて、有用な議論を継続していただいている北九州市立大学 基盤教育センターの池田 隆介先生、大阪大学 国際教育交流センターの村岡 貴子先生にも感謝を表明する。

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.

堀一成・北沢美帆・山下英里華 (2022).『ダメ例から学ぶ 実験レポートをうまくはやく書けるガイドブック』 羊土社.

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.

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.

松田寛・大村舞・浅原正幸 (2022).「UD Japanese に基づく国語研長単位解析系の構築」 言語処理学会 第 28 回年次大会発表論文集, pp. 725–730.

関連 URL

長単位形態素解析ソフトウェア『spaCy __ GSDLUW』
https://github.com/megagonlabs/UD_Japanese-GSD/releases/tag/r2.9-NE/

表 1　堀シラバスデータから spaCy_GSDLUW を用いて抽出した長単位動詞頻度表 (頻度上位 30 語まで) (堀一成 作成)

長単位動詞	長単位頻度
成る	14
為る	13
行う	13
作成する	7
書く	7
学習する	7
ディスカッションする	6
求める	5
有る	4
学ぶ	4
ことができる	4
作る	4
応じる	3
得る	3
欠席する	3
助ける	3
言う	2
纏まる	2
出す	2
利用する	2
注意する	2
心掛ける	2
止む	2
決まる	2
相談する	2
取り止める	2
富む	2
習得する	2
繋がる	2
出来る	1

表 2　堀シラバスデータから MeCab を用いて抽出した短単位動詞頻度表 (頻度上位 30 語まで) (堀一成 作成)

短単位動詞	短単位頻度
為る	108
成る	16
行う	13
居る	10
有る	9
出来る	9
行く	7
書く	7
頂く	6
求める	5
学ぶ	4
貰う	4
作る	4
答える	4
付ける	3
言う	3
応ずる	3
得る	3
助ける	3
因る	3
見る	2
纏まる	2
出す	2
心掛ける	2
決まる	2
取り止める	2
繋がる	2
探す	1
読む	1
立てる	1

表 3　堀シラバスデータから spaCy_GSDLUW を用いて抽出した長単位名詞頻度表 (頻度上位 30 語まで) (堀一成 作成)

長単位名詞	長単位頻度
題目	27
授業	23
課題	20
担当教員	15
教室対面	11
必要	10
受講者	9
受講生	9
場合	8
オンデマンド	8
曜日	8
明確	6
レポート	6
書き手	6
実際	5
説明	5
私語り	5
本授業	5
文章	5
説得力	4
論文・レポート	4
方法	4
根拠	4
身	4
技術	4
論理的	4
実習	4
５パラグラフ論証文	4
プレゼンテーション	4
分量	4

表 4　堀シラバスデータから MeCab を用いて抽出した短単位名詞頻度表 (頻度上位 30 語まで) (堀一成 作成)

短単位名詞	短単位頻度
授業	50
事	35
課題	32
題目	30
受講	23
教員	22
的	20
レポート	18
作成	18
様	17
提出	16
曜日	16
担当	16
対面	15
文章	14
教室	13
ディスカッション	13
者	12
パラグラフ	12
必要	11
実習	11
プレゼンテーション	10
学習	10
大学	9
学術	9
論文	9
情報	9
評価	9
元	9
オン	9

表 5　堀シラバスデータから spaCy_GSDLUW を用いて抽出した長単位「漢語＋する」動詞表 (全 18 語)(堀一成 作成)

作成する
学習する
欠席する
利用する
注意する
相談する
習得する
入手する
収集する
読解する
説明する
経験する
用意する
出席する
修得する
就職する
修正する
実現する