研究ノート

日本大学生産工学部研究報告 B
2023 年 6 月 第 56 巻

科学技術英語教育におけるコーパス分析の応用

小林雄一郎*，今滝暢子**

Application of Corpus Analysis in Teaching English for Science and Technology

*Yuichiro KOBAYASHI** *and Nobuko IMATAKI***

English is a *lingua franca* in the field of science and technology, and more than 90% of scientific articles are published in English. In such a situation, non-native speakers of English need to write in appropriate science English in order to publish their papers in prestigious international journals. For non-native writers, many dictionaries and vocabulary lists of science English have been created. However, they are usually intended for relatively large units of scientific disciplines, such as chemistry or mathematics. In other words, they may not provide a sufficient number of useful English expressions for the specific research topics addressed by individual researchers. To deal with the problem, the present paper introduces a methodology for collecting and analyzing a vast number of research articles on the Web, identifying the list of key expressions associated with a specific research topic or a particular section in research articles. For the convenience of the readers, free software will be used for the collection and analysis of research articles.

Keywords: English for Science and Technology, Web Scraping, Corpus Analysis

1．はじめに

科学技術の分野の実質的な公用語は英語であり，科学技術論文の 90％以上は英語で発表されている[1)]。そして，ほぼ全ての科学技術分野において，論文における適切な英語使用が求められている。従来，英語の非母語話者が英語論文を執筆する場合は，辞書の例文を参考にしたり，当該分野の論文の表現を真似たりしたあと，英語母語話者の添削を受けて，表現を修正するという手順が一般的であった。しかしながら，それらの辞書や先行研究，あるいは英文校正者のコメントが個々の研究者の研究領域に適したものであるとは限らない。

そのような状況において，コーパスと呼ばれる大規模な言語データを定量的に解析し，比較的大きい単位の研究領域で重要とされる語彙や表現のリストを作成する試みは古くから見られる。我が国で教材化された例としては，「文系共通語彙」や「理系共通語彙」を選定した『京大・学術語彙データベース－基本英単語 1110』[2)]，「学部別語彙表現」を選定した『九大英単－大学生のための英語表現ハンドブック』[3)]などを挙げられる。また，学科や研究室といった比較的小さい単位の研究領域で重要とされ

*日本大学生産工学部教養・基礎科学系准教授
**日本大学生産工学部教養・基礎科学系助教

る語彙や表現のリストを作成するための研究も存在する[4)]。

また，一口に「科学技術論文」と言っても，(a) Introduction，(b) Method，(c) Results and Discussion，(d) Conclusion といったセクションによって，典型的に用いられる語彙，表現，文法項目が異なる。その点を考慮した英語学術論文執筆支援ツールとして AWSuM[5)] などがある。このツールでは，学術分野と論文のセクション，さらに特定のセクション内における伝達内容のまとまり（ムーブ）ごとに高頻度な単語連鎖が提示される。ただし，AWSuM が執筆を支援できる学問領域が限定的であるため，直接的な恩恵を受けられる研究者や学習者は限られている。従って，任意の（比較的小さい単位の）研究領域，論文の特定のセクションにおける重要表現を特定する方法論が求められている。

そこで本稿では，ウェブ上に存在する膨大な科学技術論文を収集・分析し，特定の研究テーマに関する論文のセクションと密接に結びつく重要な表現を特定し，それらの表現の典型的な（高頻度な）使い方を把握するための方法論を紹介する。そして，読者の便を考慮し，科学技術論文の収集と分析にはフリーソフトを用いる。

2．分析データ

本稿では，AntCorGen[6)] というフリーソフトを用いて，オープンアクセスジャーナルである *PLOS ONE* から科学技術論文を自動で収集する。このツールを用いれば，Python や Ruby などでスクレイピング（ウェブからの情報収集）のコードを書く必要がなく，プログラミングに馴染みのない研究者や学生でも比較的簡単にデータの収集が可能となる。また，論文のジャンルやトピックを指定することで，特定の分野に関するデータのみを集めることができる（**図 1**）。さらに，論文の特定の部分（e.g., Introduction, Results and Discussion）のみを対象とする収集を行えば，科学技術論文のセクションごとに顕著な表現の分析ができる。

ここでは, AntCorGen を用いて，機械学習（machine learning）に関する論文の Introduction（INT）のセクションと Results and Discussion（RAD）のセクションから，1,000 本ずつテキストを収集した。**表 1** にあるように，データの総語数は約 589 万語である。

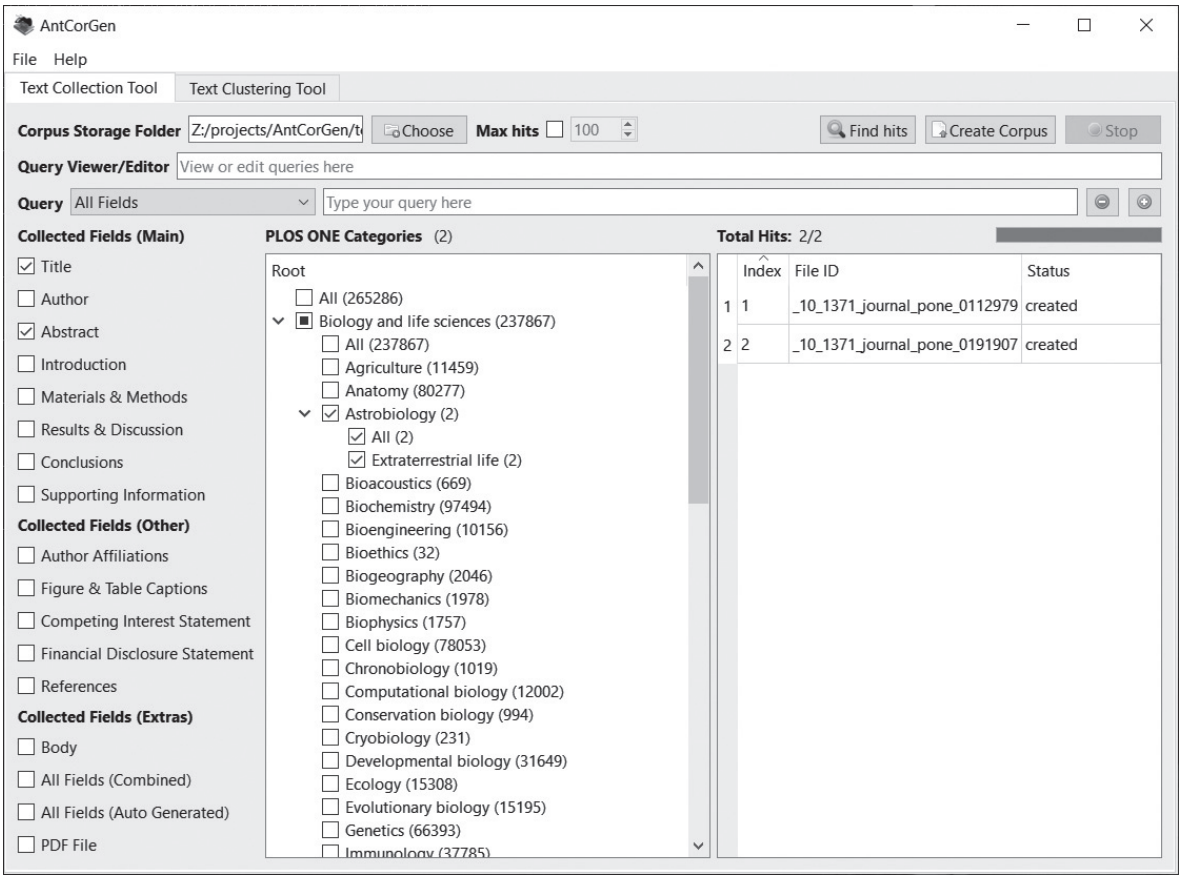

図 1　AntCorGen のスクリーンショット[注 1)]

	テキスト数	語数
Introduction（INT）	1,000	1,403,982
Results and discussion（RAD）	1,000	4,486,595
合計	2,000	5,890,577

表１　分析データの概要

３．分析の手順

本稿では，*PLOS ONE* から収集した科学技術論文における２つのセクション（INT, RAD）に特徴的な表現を定量的に抽出する。特殊目的英語（English for Specific Purposes; ESP）や学術目的英語（English for Academic Purposes; EAP）の分析では，「単語」を単位とする特徴表現抽出が行われることが多い。しかしながら，一般的に，単語の頻度はテキストの内容の影響を強く受けるため，適切な前処理や後処理を行わない限り，有益な分析結果が得られない。また，テキスト中に出現する単語の種類は数千から数万に及ぶため，頻度集計後の統計処理などでも計算機に大きな負荷をかける。そこで，以下の分析例では，コーパス言語学で古くから使われている Biber（1988）[7]の 67 種類の言語項目を特徴表現抽出の単位とする。Biber の言語項目を用いることで，語彙，品詞，統語，談話というテキストの様々な層を分析することが可能になる。67 種類の言語項目の頻度を求めるにあたっては，Multidimensional Analysis Tagger[8]というフリーソフトを使用する。そして，Multidimensional Analysis Tagger で各言語項目の頻度を計算したのち，フリーの統計処理ツールである R[9]を用いて，Wilcoxon の順位和検定に基づく特徴表現抽出[10]を行う。

４．結果と考察

表２は，INT と RAD における 67 種類の言語項目の相対頻度（100 語あたり）に対して Wilcoxon の順位和検定に基づく特徴表現抽出を行なった結果（上位 10 位まで）である[注2)]。なお，表中の keyness はセクション間での中央値の差の大きさを表している。そして，**図２**は，上位 10 位までの言語項目の相対頻度を箱ひげ図で可視化したものである。

表２および図２を見ると，上位 10 位までの全ての言語項目において，INT の方が RAD よりも中央値が高いことが分かる。それでは，これら上位の言語項目は，科学技術論文の中で，実際にどのように使われているのだろうか。この点を明らかにするために，AntCorGen と同じ開発者が公開している AntConc[11]というフリーソフト（**図３**）を用いて，頻度１位の TO（infinitives）の使用例を調査した。

言語項目	keyness	中央値	
		INT	RAD
TO	855618.0	1.51	0.89
PEAS	854302.0	0.47	0.15
AWL	800615.5	5.50	5.19
VPRT	779641.0	4.91	3.47
TTR	744326.0	73.0	67.0
JJ	742262.5	10.53	8.84
NOMZ	725703.5	5.07	3.97
SPAU	722426.5	0.42	0.27
CONJ	704113.5	0.82	0.58
SUAV	700275.0	0.44	0.27

表２　Wilcoxon の順位和検定に基づく特徴表現抽出の結果（上位 10 位まで）

INT における TO の使用例を AntConc の Cluster 機能で分析したところ，［to＋動詞］の頻度１位は［to be］であり，頻度２位以降に，［to predict］，［to identify］，［to detect］，［to find］，［to classify］などが続く。これらの表現は，ここで分析対象としている学問領域，すなわち，多様な特徴量に基づき何らかの統計的予測・分類を行う機械学習の目的を反映している。そして，一例として，頻度２位の［predict］に注目し，［predict ＋名詞句］を調べると，［predict the future price(s)］，［predict the stock market］，［predict essential genes］などの高頻度パターンを発見できる。また，それ以外の表現では，［predict whether］や［predict accurately］などが頻出している。このような分析は，言語学や言語教育のための有益な知見を与える。また，このような高頻度パターンを学習者自身に発見させる Data Driven Learning（DDL）で活用することもできる[12]。言うまでもなく，名詞や動詞といった内容語の頻度はコーパスに収録されているテキストの影響を受けるが，そうであるからこそ，任意の研究領域における重要表現を抽出することが可能になる。

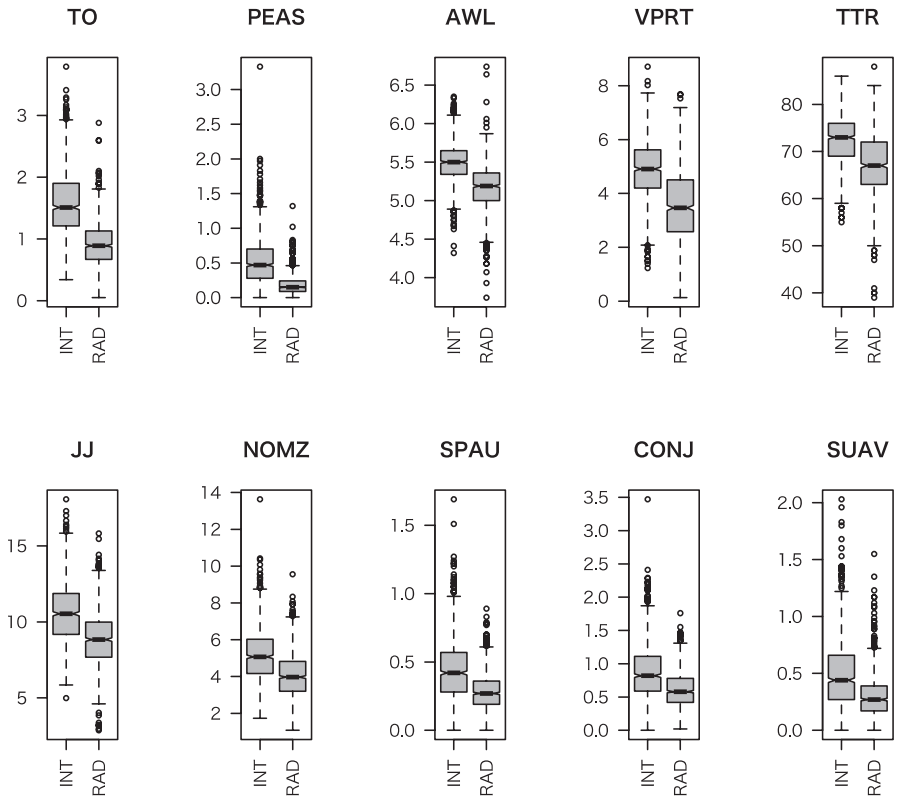

図2　上位 10 位までの言語項目の箱ひげ図

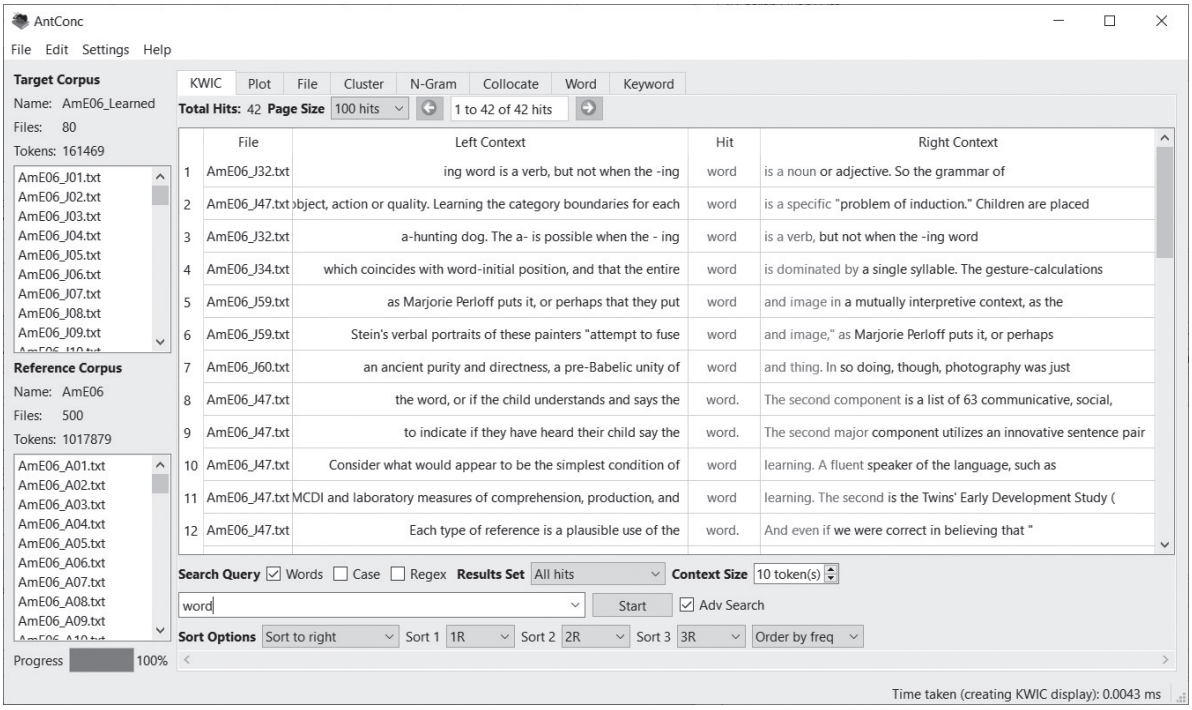

図3　AntConc のスクリーンショット[注3)]

5．おわりに

ここまで，*PLOS ONE* から科学技術論文を自動収集し，特定の研究テーマや論文のセクションにおける高頻度表現を抽出するための方法を紹介してきた。さらなる方向性として，*n*-gram 分析や構文解析の結果に基づく特徴表現抽出を行うことで，テキストの統語的な特徴を詳細に検討することができる[13)]。また，科学技術分野の英語として質の高い論文と質の低い論文の比較，あるいは母語話者が書いた論文と日本人が書いた論文の比較を行うことで，日本人には習得の難しい言語表現を明らかにすることができる[14)]。

注

1　https://www.laurenceanthony.net/software/antcorgen/

2　個々の言語項目については，Multidimensional Analysis Taggerのマニュアルなどを参照。https://sites.google.com/site/multidimensionaltagger

3　https://www.laurenceanthony.net/software/antconc/

引用文献・使用ツール

1）Montgomery, S. L.（2013）. *Does science need a global language? English and the future of research.* Chicago: University of Chicago Press.

2）京都大学英語学術語彙研究グループ＋研究社（2009）.『京大・学術語彙データベース―基本英単語1110』東京：研究社.

3）九州大学英語表現ハンドブック編集委員会（2014）.『九大英単―大学生のための英語表現ハンドブック』東京：研究社.

4）田中省作・冨浦洋一・徳見道夫（2014）.「機関リポジトリから得られる著者の語彙分布に基づいた部局別重要語彙の選定」『じんもんこん2014論文集』*3*，207-212.

5）Mizumoto, A.（2017）. Initial evaluation of AWSuM: A pilot study. *Vocabulary Learning and Instruction, 6*(2), 46-51.

6）Anthony, L.（2022）. AntCorGen（Version 1.2.0）. Waseda University. https://www.laurenceanthony.net/software

7）Biber, D.（1988）. *Variation across speech and writing.* Cambridge: Cambridge University Press.

8）Nini, A.（2019）. The multi-dimensional analysis tagger. In T. B. Sardinha & M. V. Pinto（Eds.）, *Multi-dimensional analysis: Research methods and current issues*（pp.67-94）. London: Bloomsbury Academic.

9）R Core Team（2021）. R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.r-project.org/

10）Paquot, M., & Bestgen, Y.（2009）. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In A. Jucker, D. Schreier, & M. Hundt（Eds.）, *Corpora: Pragmatics and discourse*（pp.247-269）. Amsterdam: Rodopi.

11）Anthony, L.（2022）. AntConc（Version 4.1.2）. Tokyo, Waseda University. https://www.laurenceanthony.net/software

12）Lenko-szymanska, A., & Boulton, A.（Eds.）（2015）. *Multiple affordances of language corpora for data-driven learning.* Amsterdam: John Benjamins.

13）小林雄一郎・田中省作・冨浦洋一（2012）.「N-gramを素性とするパターン認識を用いた英語科学論文の質判定」『情報処理学会研究報告』*2012-NL-205*，1-6.

14）小林雄一郎・田中省作（2014）.「メタ談話標識を素性とするランダムフォレストによる英語科学論文の質判定」岸江信介・田畑智司（編）『テキストマイニングによる言語研究』（pp.137-151）. 東京：ひつじ書房.

（R５．２．10受理）