BCCWJ コアデータの頻度情報に基づく 日本語論文・レポートライティング指導の試み

堀 一成（大阪大学　全学教育推進機構）†
坂尻 彰宏（大阪大学　全学教育推進機構）†

Attempt to Provide Educational Support for Japanese Academic Writing, Using Frequency Information Retrieved from the BCCWJ Core Corpora Data

Kazunari Hori (Osaka University, Center for Education in Liberal Arts and Sciences)
Akihiro Sakajiri (Osaka University, Center for Education in Liberal Arts and Sciences)

１．概要

論文・レポートのライティング指導の基礎データとするため、BCCWJ・コアデータ「『代表性を有する大規模日本語書き言葉コーパスの構築：２１世紀の日本語研究の基盤整備』総括班(2011)」より語彙頻度情報をマイニングし、指導に活用した事例を報告する。

論文・レポートの書き方指導書などでは、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・文例の根拠が明示されていることはまれである。我々は、BCCWJ を基礎とすることで、特定の著者や学会に偏らないデータが得られ、その成果をライティング指導に活用することで、より汎用性の高いレポート作成技能を受講者に身につけさせることができると考えた。

今回は、このような試みの第一報として報告する。BCCWJ コアデータの白書データを選び、動詞・名詞の頻度情報を得た。一般文でも利用される頻度が高いと判断される語を、頻度上位のリストから除き、論文・レポートで用いることを推奨する用語集として受講者に提供した。実際のセミナー授業での活用の様子なども併せて報告する。

２．文章指導における言語資源活用事例と本研究の目的

これまで発行されたライティング関連書籍や教材には、少ないながらも、アカデミックな文章に使われる表現例や文例を提示し、参考にさせる優れたものがある。たとえば「二通　他(2009)」は、実際の学術論文から文例をとり、用いるべき表現として紹介している。しかしその表現が採用された根拠（一般文と異なり学術的文章でより用いられやすいとする計量的根拠）は提示されていない。また BCCWJ などのコーパスデータに基づく Web 日本語作文支援システム「なつめ」「仁科（2012)」は、入力した語に対する共起情報を例文根拠情報と共に表示し、用いると良い表現を知ることができる。しかし最初にシステムに入力すべき語（表現）の知識がなければ有効に用いることが難しい。

本研究では、特に大学初年次学生のアカデミックな表現に対する知識不足に対応するための教材を開発し、かつその教材が、教員・指導者の経験や内省によるものでなく、コーパスなどの根拠情報から定量的に得られるものとすることを目的としている。それにより、受講者の教材に対する納得感を向上させたいと考えている。

† hori@celas.osaka-u.ac.jp,　sakajiri@celas.osaka-u.ac.jp

今回の発表では、このような教材開発の最初の試みとして、BCCWJ コアデータより動詞・名詞の頻度情報を抽出し、アカデミックな文に特徴的に使われると考えられる用語のリストを作成し、大学学部初年次のライティングセミナー授業で教材として活用した事例の報告を行う。

最近では、英語ライティングのテキスト「富岡(2012)」が発行されているが、これは上記の考えに基づくものである。British National Corpus に基づき、使用頻度順に 150 の動詞・助動詞を並べ使用例文と解説文を提示したものである。また「古泉　他(2011)」では、頻度情報を教育に応用した事例が紹介されているが、単語そのものを覚えることを主眼としたものであった。

３．頻度リストの作成方法

以下に教材として提示した動詞・名詞の頻度情報を作成した手順を説明する。

（１）BCCWJ コアデータ 白書データからの情報抽出

まず、BCCWJ コアデータのうち、CSV 形式のものを Excel2010 に読み込み、各種フィルタリング処理を行った。まず、比較的長い特徴的な単語を抽出するため、長単位情報を基に選択する。品詞情報が「動詞－一般」あるいは「名詞－普通名詞－一般」となっているもののみをそれぞれ抽出した。その単語リストの出現頻度を統計ツールで計算し、頻度順に並べ替えた。

（２）一般文でも利用される頻度が高いと判断される語のフィルタリング

『日本語教育のための基本語彙調査』「国立国語研究所(2001)」に掲載されている語彙のうち、「より基本的な語」とされた約２０００語を、除去参照データとした。２０００語のうち動詞と分類される語、および一般名詞と分類されている語のリストを作成し、（１）で説明した頻度順リストから除く処理をおこなった。

（３）人手による用語選定と整形

上記のように機械的操作によって得られたリストには、大学生のアカデミックライティングにあまり用いることのない単語も含まれているので（元データが白書であるため)、最後に報告者(両名)が実際のライティング指導資料として適当と判断する語に絞り、用語表として学習者に提供した。動詞は上位３００語、名詞は上位１７６語のリストとなっている。

４．作成データのライティング指導への活用

作成した頻度データを、報告者（坂尻）が担当するライティング指導セミナー授業で教材として提供した。受講者には、ライティングの実践において口語的な表現を避けるための一つの方法として、あるいは、表現に迷った際の判断基準の一つとして、使うことを勧めた。また、とりあえずの使い方として、使いたい表現や迷う表現を 50 音で検索して頻度を確認し、国語研の Web ツール（少納言と NLB）を使って文脈での用法や出典（ブログ等か書籍等か）を参照することを提案してみた。

まず、登録等の必要が無い国語研の BCCWJ 検索システム「少納言」を紹介した。配布資料で紹介した用語を利用するに際して、どのような文脈中でその語が使われているかを少納言で検索し、例をよく読んで納得してから使うべきだと指導した。

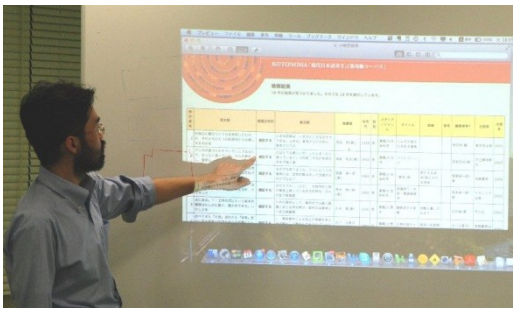

図 1 報告者（坂尻）が少納言の利用方法を担当セミナーで説明している場面

図 2　配布単語頻度データと少納言の説明を聞き、利用方法を学んでいる受講学生

また、同様の用例検索ツールとして、NINJAL-LWP for BCCWJ（以下、NLB）「Pardeshi, 赤瀬川(2012)」も紹介した。NLB は、国立国語研究所のプラシャント・パルデシ氏と Lago 言語研究所の赤瀬川史朗氏が中心になって開発した BCCWJ オンライン検索システムである。NLB はコンコーダンサとは異なるレキシカルプロファイリング手法を用いたコーパス検索ツールで、名詞や動詞などの内容語の共起関係や文法的振る舞いを網羅的に表示できるのが最大の特長とされている。受講者には、用例を調べようとする語について、文法項目で分けられた共起情報が細かく検索できるため、より適切な表現を見つけることができると説明した。

図 3　報告者（坂尻）が NLB の操作法を説明している場面

本資料に基づくライティング授業に参加した受講学生（全員学部１年生）から意見を徴収した。７割ほどの受講生は、「(資料をライティングの際の参考に）使えそう」「使ってみたい」との意見であったが、「分からない」「使えなさそう」との意見もあった。さらに、「使い方や応用のポイントをより詳しく指導してほしい」あるいは「コーパスサイトとの併用の仕方を詳しく説明してほしい」などの意見もあったので、受講者のうち希望する者にはコーパスサイトを開いてのインストラクションを行った。その結果、説明を受けた学生は、納得しているようであった。

５．今後の展開

本報告は、BCCWJ データを有効活用するための手法開発の試行という位置づけである。今回の試行した手順をきっかけに、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。

◎　対象コーパスデータの検討

今回の抽出は、試行としての時間的制約により、抽出対象を白書データのみとした。この抽出対象は語彙に偏りがあるなど、十分でないと認識している。まず、BCCWJ DVD データを本稿執筆時点で入手手続き中であり、入手後すみやかに対象データを DVD データとしたい。またアカデミックな文章に比較的近い表現（硬い文章）が多く含まれるであろうと判断し白書データジャンルを選定したが、BCCWJ の図書館データや教科書データの内、対象とすべきデータはまだ多数あると認識しており、適切な対象を追加選定したい。さらに BCCWJ だけでなく、Wikipedia 項目の内、学術的文章の参考になりうると判定できるものについては、対象としたいと考えている。

◎　特徴的な語・表現の抽出方法の改良

今回、語の抽出方法は、得られたリストから基本的な２０００語に含まれるものを除くという、簡易な手法であった。今後適切なデータ集団の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。

◎　作業手順のプログラム化

今回の頻度情報資料は Microsoft Excel を用い、手作業で抽出やフィルタリングを行った。今後作業対象コーパスの拡大を予定しており、できる限り早急に作業をプログラム化したいと考えている。

◎　資料インストラクション手法の改善

受講生に資料の有効活用法を説明する方法も改善が必要である。前述したとおり、受講生より、資料の活用法が良くわからないとの感想も得られている。より時間を掛け、頻度リストや関連 Web ツールを使用して、より良い文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。そのためのわかりやすい教材作りも必要だと考えている。

６．まとめ

BCCWJ・コアデータより語彙頻度情報をマイニングし、論文・レポートのライティング指導に活用した事例を報告した。BCCWJ コアデータの白書の長単位情報を選び、動詞・名詞の頻度情報を得た。頻度上位のリストから一般文でも利用される頻度が高い語を抜き、用いることを推奨する用語集として受講者に提供した。

謝　辞

本研究は、文部科学省　科学研究費補助金　基盤研究（B）課題番号:22320103「多言語会話文・語彙データベース構築と異文化交流におけるその活用に関する研究」（研究代表者：萬宮健策）による補助を得ている。

文　献

古泉隆、梁志鋭、阪上辰也、坂東貴夫、天野修一、新實葉子(2011)「日本語複合動詞を学ぶための Web 教材開発　－BCCWJ の頻度データに基づいて－」特定領域研究「日本語コーパス」平成２２年度公開ワークショップ（研究成果報告会）予稿集、pp.27-32

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

「代表性を有する大規模日本語書き言葉コーパスの構築：２１世紀の日本語研究の基盤整備」総括班(2011)「特定領域研究『日本語コーパス』研究成果報告」

富岡龍明(2012)「コーパス活用　英語基本語を使いこなす　動詞・助動詞編」研究社

仁科喜久子　監修(2012)「日本語学習支援の構築」凡人社

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

Pardeshi, Prashant、赤瀬川史朗(2012)「コーパスを利用した基本動詞ハンドブック作成　－コーパスブラウジングツール NINJAL-LWP の特徴と機能－」言語処理学会第１８回年次大会　予稿集 pp.575-578

関連 URL

国立国語研究所の言語コーパス整備計画 KOTONOHA　http://www.ninjal.ac.jp/kotonoha/

国立国語研究所 BCCWJ 検索ツール「少納言」　http://www.kotonoha.gr.jp/shonagon/

NINJAL-LWP for BCCWJ ホームページ　http://ninjal-lwp-bccwj.ninjal.ac.jp/

東京工業大学「なつめ」ホームページ　http://hinoki.ryu.titech.ac.jp/natsume/

表 1　教材とした動詞頻度表の一部

国立国語研究所　書き言葉コーパス（BCCWJ コアデータ）より抽出した				
論文・レポートに役立つ動詞表現集（頻度順）				
２０１２年７月１３日　大阪大学　堀　一成、坂尻　彰宏				
動詞リスト番号	頻度	動詞	読み	頻度順位
1	779	する	スル	1
2	321	図る	ハカル	5
3	282	実施する	ジッシスル	6
4	233	推進する	スイシンスル	8
5	129	基づく	モトヅク	10
6	95	応ずる	オウズル	15
7	92	踏まえる	フマエル	16
8	91	増加する	ゾウカスル	17
9	86	活用する	カツヨウスル	19
10	83	利用する	リヨウスル	20
11	79	取り組む	トリクム	24
12	78	伴う	トモナウ	25
13	76	有する	ユウスル	28

表 2　教材とした名詞頻度表の一部

国立国語研究所　書き言葉コーパス（BCCWJ コアデータ）より抽出した				
論文・レポートに役立つ名詞表現集（頻度順）				
２０１２年７月１３日　大阪大学　堀　一成、坂尻　彰宏				
名詞リスト番号	頻度	名詞	読み	頻度順位
1	296	整備	セイビ	3
2	210	推進	スイシン	8
3	170	地域	チイキ	10
4	132	障害	ショウガイ	14
5	129	注	チュウ	16
6	126	図表	ズヒョウ	17
7	106	情報	ジョウホウ	23
8	99	支援	シエン	26
9	98	実施	ジッシ	27
10	93	資料	シリョウ	29
11	92	対象	タイショウ	30
12	89	推移	スイイ	33