BCCWJ 教科書データより抽出した頻度情報に基づく日本語ライティング指導教材の作成

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）†
石島　悌（大阪府立産業技術総合研究所）‡

Creation of Teaching Materials for Japanese Academic Writing, Using Frequency Information Retrieved from the BCCWJ School Text 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)
Dai Ishijima (Technology Research Institute of Osaka Prefecture)

１．概要

大学学部初年次生向け論文・レポートのライティング指導の基礎データとするため、国立国語研究所が開発した「現代日本語書き言葉均衡コーパス（略称:BCCWJ）」の高校教科書データ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、より広範囲に応用できる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。

本稿は、このような試みの第二報である。第２回コーパス日本語学ワークショップにおいては、試行との位置づけで、BCCWJ コアデータの白書データに基づく成果を報告した「堀, 坂尻(2012)」。今回は BCCWJ　DVD 版公開データの特定目的サブコーパス教科書（コーパス記号 OT）のうち高校教科書とラベル付けされたものを対象とし、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。また作業を MySQL 上での SQL プログラム実行で行い、一部の自動化を実現した。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

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

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、特に大学学部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、教材の客観性を高めるとともに、受講者からの信頼性をより向上させたいと考えている。

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 教科書（OT）高校限定データからの情報抽出

まず、BCCWJ 教科書(OT)データを CSV 形式として MySQL 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こととし、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 SQL コマンドで計算し、頻度順に並べ替えた。作業の詳細については付録で説明する。

（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（理科や音楽の用語など、各教科でのみ使われ一般用語として紹介することが適当でないと判断した）、最後に報告者(堀・坂尻)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 330 語、名詞は上位 312 語のリストとなっている。活用法を多様にするため、頻度上位から順に紹介するリストと、その内容を読みの五十音順に並べ替えたリストの両方を提供した。

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 2013 年度ライティング指導セミナー授業で教材として提供した。

４．１　受講者への説明

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、前述のリストの使用を勧めた。レポート作成の際に、ことば遣いに迷った場合、たとえば、五十音で表を検索してそのことばがあれば、頻度の高い使用可能なことばであることわかる。もし、表になければ、口語的表現、堅すぎる表現、たまたま一覧に無いかのいずれかと判断されるので、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授業で説明している場面

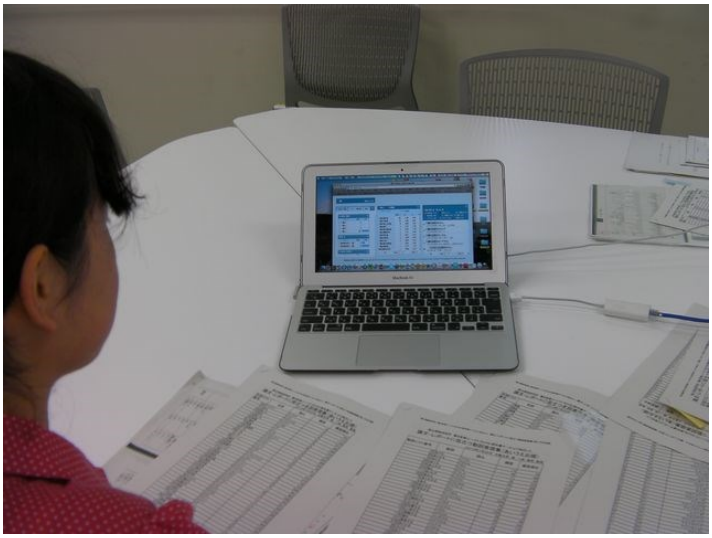

図 2　配布単語頻度データを元に NLB の利用方法を学んでいる受講学生

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。

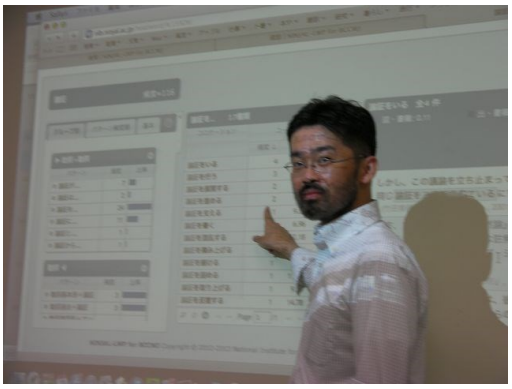

図 3　報告者（坂尻）が NLB の操作法を説明している場面

４．２　受講者の反応

（１）頻度別の一覧を見て

受講者は、論文・レポートに使用することばが、彼らにとって、必ずしも未知で難解なことばではないことに気づき、レポート作成の際に、「ことば遣い」に臆する必要のないことに納得していた。

（２）五十音順の一覧を見て

授業で各自のレポートを見ながら、気になることば遣いを表やサイトで検索し、確認させると学生の理解の度合いも高かった。

総じて一覧表に対する学生の反応は「おもしろい」「参考になる」と良好であったが、表の意味やサイトとの連携などについては細かい指導や解説が必要で、実際に使用する場合には工夫が必要である。

５．今後の展開

本報告は、BCCWJ データを教育に有効活用するための手法開発という位置づけである。今回の手順をきっかけに、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。

◎　解析対象コーパスデータの広範化

今回の抽出は、BCCWJ データのうち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高校教科書データを解析対象としたが、BCCWJ の図書館データや書籍データの内、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と考えている。さらに BCCWJ だけでなく、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、広く対象としたいと考えている。近年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や、国立情報学研究所の CiNii 論文情報、Wikipedia の学術的項目説明文などを対象にすべきと考えている。

◎　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、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。

◎　作業手順のさらなるプログラム化

今回は BCCWJ の一部データから品詞別頻度情報を得る作業を SQL プログラム化した。それ以外の資料整形等作業は Microsoft Excel を用い、手作業で行った。今後作業対象コーパ

スの拡大を予定しており、R,Mecab なども利用し、広範囲な作業をプログラム化したいと考えている。また開発したプログラムを広く利用してもらえるよう公開する予定である。

◎　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、作成した資料のみを渡すだけでは、有効活用は期待できない。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も早急に整備したいと考えている。

６．まとめ

BCCWJ 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 高校教科書データ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、併せて BCCWJ 活用 Web システムの活用紹介も行った。

謝　辞

本研究は、学術研究助成基金助成金 挑戦的萌芽研究 課題番号: 25540163「XML コーパスからの抽出データに基づく日本語学術ライティング教材作成法の研究」（研究代表者：堀一成）による補助を得ている。

文　献

堀一成、坂尻彰宏(2012)「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」　第２回コーパス日本語学ワークショップ予稿集、pp.1-6

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」

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　コーパス開発センター　KOTONOHA 計画
http://www.ninjal.ac.jp/corpus_center/kotonoha.html

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 (NLB) ホームページ　http://nlb.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/

付録　BCCWJ データからの頻度情報抽出処理の詳細

以下に、３．頻度リストの作成方法（１）で略述した、BCCWJ データを Windows 上の MySQL に搭載し、頻度情報を抽出するための作業手順を詳述する。実際はステップごとに出力をチェックしつつ作業を進めたが、ここでは根幹となる作業内容のみを記す。

作業した環境は、OS: Windows 8 Enterprise 64Bit、MySQL バージョン:5.6.11 for Win64bit、Excel 2013 Professional for Win64bit である。

まず、BCCWJ　DVD 版公開データ Disk2 の LUW フォルダ下 OT フォルダ内の OT.ZIP を展開する。OT.txt が得られる。これを Excel で読み込み、作業 Index 用に index_id 列を作り、１から順に最後まで連番（最後の行は 924835）を振る。表中の数字データで使われている","を取り除き、文字コードを UTF-8 BOM なしにして、BCCWJ_LUW_OT.csv というファイル名で CSV ファイルとして保存した。

次に MySQL 上でデータを読み込むためのテーブル bccwjluwot を作成する(SQL 省略)。テーブル bccwjluwot にデータを読み込ませる SQL 文は以下のとおりである。

```
LOAD DATA INFILE 'BCCWJ_LUW_OT.csv' INTO TABLE bccwjluwot FIELDS TERMINATED BY ',';
```

以下動詞頻度表の場合のみを紹介する。高校教科書（判定は、articleID の 4 文字目が '3' かどうか）でかつ品詞が「動詞 - 一般」のデータのみを Select し、新しく VIEW とする。

```
CREATE VIEW highverblist AS SELECT index_id, l_orthBase, l_formBase FROM bccwjluwot
WHERE substring(articleID,4,1)='3' and l_pos = '動詞--一般';
```

これで highverblist に高校教科書で使われている一般動詞の長単位情報が抜き出せている。さらに動詞頻度順表をつくる。

```
SELECT l_orthBase,l_formBase, COUNT(*) INTO OUTFILE 'orderedhighverblist.csv' FIELDS
TERMINATED BY ',' FROM highverblist GROUP BY l_orthBase ORDER BY COUNT(*) DESC;
```

これで、動詞頻度順データが orderedhighverblist.csv に保存できた。

Query OK, 5978 rows affected (0.13 sec)

とのメッセージから頻度順に高校教科書使用動詞が 6000 件弱、抽出されたことがわかる。

表 1　教材とした動詞頻度表（出現頻度順）の一部

国立国語研究所　書き言葉コーパス（BCCWJ 高校教科書データ）より抽出した				
論文・レポートに役立つ動詞表現集（出現頻度順）				
２０１３年７月２２日　大阪大学　堀　一成、坂尻　彰宏				
動詞リスト番号	**動詞**	**読み**	**頻度**	**頻度順位**
1	する	スル	3447	1
2	なる	ナル	3204	2
3	ある	アル	2320	3
4	いう	イウ	1356	4
5	できる	デキル	676	5
6	もつ	モツ	670	6
7	つくる	ツクル	529	7
8	わかる	ワカル	380	8
9	利用する	リヨウスル	355	9
10	みる	ミル	309	10
11	異なる	コトナル	236	13
12	変化する	ヘンカスル	223	14
13	まとめる	マトメル	168	20

表 2　教材とした動詞頻度表（五十音順）の一部

国立国語研究所　書き言葉コーパス（BCCWJ 高校教科書データ）より抽出した				
論文・レポートに役立つ動詞表現集（五十音順）				
２０１３年７月２２日　大阪大学　堀　一成、坂尻　彰宏				
動詞リスト番号	**動詞**	**読み**	**頻度**	**頻度順位**
26	あげる	アゲル	112	35
90	あたえる	アタエル	48	109
52	あたる	アタル	71	65
249	あてる	アテル	18	375
228	あふれる	アフレル	20	328
324	誤る	アヤマル	11	590
111	あらわす	アラワス	39	138
329	表せる	アラワセル	11	598
75	あらわれる	アラワレル	54	91
3	ある	アル	2320	3
100	あわせる	アワセル	42	124
106	安定する	アンテイスル	40	132
4	いう	イウ	1356	4

表 3　教材とした名詞頻度表（出現頻度順）の一部

国立国語研究所　書き言葉コーパス（BCCWJ 高校教科書データ）より抽出した				
論文・レポートに役立つ名詞表現集（出現頻度順）				
２０１３年７月２２日　大阪大学　堀　一成、坂尻　彰宏				
名詞リスト番号	名詞	読み	頻度	頻度順位
1	情報	ジョウホウ	384	5
2	物体	ブッタイ	296	8
3	課題	カダイ	176	28
4	しくみ	シクミ	132	40
5	動き	ウゴキ	124	44
6	役割	ヤクワリ	115	49
7	構造	コウゾウ	107	52
8	一定	イッテイ	101	56
9	現象	ゲンショウ	99	58
10	考え方	カンガエカタ	97	60
11	仮説	カセツ	93	63
12	課程	カテイ	85	70
13	手順	テジュン	84	71

表 4　教材とした名詞頻度表（五十音順）の一部

国立国語研究所　書き言葉コーパス（BCCWJ 高校教科書データ）より抽出した				
論文・レポートに役立つ名詞表現集（五十音順）				
２０１３年７月２２日　大阪大学　堀　一成、坂尻　彰宏				
名詞リスト番号	名詞	読み	頻度	頻度順位
294	アイデア	アイデア	10	1781
191	歩み	アユミ	14	1112
200	表し方	アラワシカタ	14	1153
172	安定	アンテイ	16	1012
148	言い方	イイカタ	18	860
185	意義	イギ	15	1089
281	維持	イジ	10	1729
72	意識	イシキ	30	376
179	位置関係	イチカンケイ	15	1047
150	一切	イッサイ	18	869
298	一体	イッタイ	10	1794
8	一定	イッテイ	101	56
80	イメージ	イメージ	29	411