BCCWJ を用いた語彙・文法情報のプロファイリングとその応用

千葉庄寿（日本語教育班連携研究者：麗澤大学外国語学部）†

Doing Lexical and Grammatical Profiling with BCCWJ

Shoju Chiba (Faculty of Foreign Studies, Reitaku University)

１．BCCWJ をもちいた語彙・文法情報の評価

英語コーパス言語学の初期の展開において，辞書学をはじめとする語彙研究への関心が重要な役割を果たしたことが知られている(Biber *et al*.1998)。日本の英語教育においても，基本語リストの作成(大学英語教育学会基本語改訂委員会編 2003)や英和辞典の編纂などに語彙教育への大規模コーパスの応用事例をみることができる。

2011 年に公開される予定の『現代日本語書き言葉均衡コーパス』(BCCWJ)は，サンプリング手法を用いて収録するサンプルに(少なくとも部分的に)統計的な代表性をもたせた大規模な「均衡コーパス」である(前川 2007:14; 丸山 2009:129)。このような「書き言葉のサンプル」たる設計思想をもつコーパスは「サンプルコーパス」(sample corpus, 齊藤ほか 2005[2]: 23)とも呼ばれ，後藤(2003: 8-9)が述べる，言語研究用に設計された「最狭義のコーパス」の最右翼の候補として，日本語の研究において未だ立ち後れている大規模コーパスを活用した定量的な語彙研究に画期的な活路を開くことが期待できる。

具体的には 2 つの活用方法が考えられよう。第一に,「書き言葉のサンプル」である BCCWJ そのものを分析し，さまざまな場面・用途に応用できる語彙データを得ることができる。実際に，BCCWJ の応用と評価を目的とした BCCWJ の研究班(研究項目 B01)のいくつかが BCCWJ の定量的な語彙研究に取り組んでいる。例えば，言語政策班は「国語政策や国語教育に役立つさまざまな語彙表を作成していくための基盤として，分野ごとの特徴度の設定と，頻度に基づく語彙レベルの設定，という二つの作業を行う」(田中 2009: 666)ことをその主要な任務としている (*cf.* 前川 2006: 1-2)。

一方，BCCWJ の利用価値は BCCWJ そのものの語彙の研究にとどまるものではない。「書き言葉のサンプル」としての BCCWJ との比較を通じ，他のコーパスデータの語彙特徴を測ることもできる。このようなコーパス間の比較の手法は BCCWJ のプロジェクトでも議論されている。近藤(2008)は，対数尤度比を指標として用い，形態素解析されたデータを用いた語彙を計量した特徴語抽出の手法を用いて教科書の語彙特徴を分析している。同様の手法を用いて，日本語教育班でも BCCWJ に基づいた日本語教育のための語彙リストの作成を試みている(橋本ほか 2008; 山内編 2008)が，こちらは BCCWJ の書籍データを話題の内容に応じた小規模なサブセットに分割し，個々の話題データの語彙特徴を抽出するものである。

本稿が射程とするのは後者であり，BCCWJ を短単位辞書 UniDic (伝ほか 2007)を用いて解析し作成した語彙情報データベースに基づき，BCCWJ の語彙・文法情報と他のコーパス(テクスト)の語彙・文法情報との比較を手軽に行うシステムの構築を報告する(本システムの公開情報については論文末を参照)。本ポスターではまた，BCCWJ の語彙情報データベースを利用した語彙・文法情報の分析ツールを用い，日本語教育における教材の開発と評価への

† schiba@reitaku-u.ac.jp

活用事例を紹介するとともに，現在の課題と今後の展望を述べる。さらに，BCCWJ によるテクストの評価についてのより広範な応用の可能性についても議論し，語彙・文法に関する信頼できる量的情報を将来どのように活用できるかを模索したい。

日本語教材に語彙情報を付与する試みとして，これまで「リーディング チュウ太」(川村 2000)や「あすなろ」(仁科 2000)などの優れた日本語読解学習支援システムが構築されてきている(各サービスの URL は論文末を参照)。しかし，これらいずれも教材テクストの分析による語彙頻度や「日本語能力試験」の語彙レベルなどの情報は考慮するものの，大規模サンプルコーパスの語彙・文法情報を活用するには至っていない。また，コロケーション情報に基づく語彙分析に利用できるオンラインツールとして日本語用例・コロケーション抽出システム「茶漉」(深田 2007)があるが，教材データなど自前のデータの分析目的に簡便に利用することはできない。

２．語彙情報データベースと語彙・文法情報分析ツール

本稿で構築した BCCWJ の語彙情報データベースは，動作が軽く，インストールおよびデータベースファイルの扱いが簡単なパブリック・ドメイン [1] の関係データベースエンジン(RDBMS)である SQLite 3.7.x で構築する。

語彙情報データベースは簡便を期し，単独出現する短単位の語彙素のテーブルと 2 グラム bigram の頻度に関するテーブルの 2 種類について BCCWJ の語彙情報を収録している。前者については UniDic の短単位の語彙素と品詞のペアを「レマ」lemma (「レンマ」とも)としてインデックスを作成した。後者は隣り合う 2 つの短単位のレマのペアについてインデックスを作成している。

分析にあたっては，分析対象のデータを UniDic で事前に解析する必要がある。本システムの利用に際しては，Windows 環境で手軽に利用できる UniDic の解析フロントエンドである「茶まめ」を使って分析対象のファイルを解析し，結果をファイルに出力しておく。BCCWJ の語彙情報データベースと分析データの解析に同じ解析環境(同一バージョン，同一環境設定の UniDic)を使うことにより，出力結果をシームレスに対応させ，齟齬なく評価することができる。

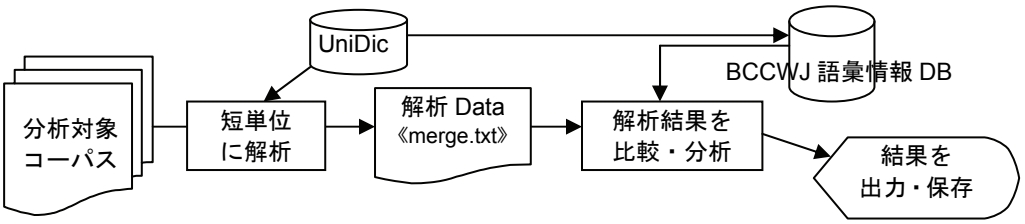

図 語彙・文法情報分析システムの概略

BCCWJ の語彙情報データベースを使い，分析対象となる語彙・文法情報を分析するツールをオフライン用のスクリプトと Web サーバ上で動作する CGI システムとしてそれぞれ Perl で構築した。Perl は ActivePerl (製品 URL は論文末を参照)の 5.8 以降のバージョンであれば，標準 [2] で SQLite が利用できるようになっている。

---

[1] http://www.sqlite.org/copyright.html

[2] 通常，Perl は CPAN (Comprehensive Perl Archive Network) を通じてモジュールを入手す

語彙・文法情報分析ツールが実装している分析手法と指標は以下の 3 種類である。

1. 語彙頻度：分析対象に現れた短単位の語彙素と品詞のペア(レマ)について BCCWJ と分析対象のコーパスの頻度を検索し，両者の数値を対数尤度比(Log-Likelihood Ratio，以下 LLR, *cf*. Kilgarriff 2001; 近藤 2008)で比較する。
2. 2 グラムの頻度：隣り合う 2 つの短単位の基本形と品詞のペアについて LLR で比較する。
3. コロケーションの計量：隣り合う 2 つの短単位の基本形と品詞のペアについて，各短単位の出現頻度と共起頻度を元に MI スコア，*t* スコアを算出し，比較する。

現在のバージョンではデータベースのサイズの問題で活用型情報による分析は行わず，語彙素情報と品詞情報のみを扱っている。

分析するコーパスが複数のファイルからなる場合には，「茶まめ」を使い解析結果を単一ファイルに出力(merge)しておくことにより，各分析ツールの結果に分析対象のコーパスデータの文書数をもとに各語彙情報の出現割合を出力する。これにより，例えば，専門用語の偏りなど，該当する用語がどの程度偏って出現しているかどうかを確認できる。

さらに，上記分析ツールは BCCWJ 全体の頻度に加え，BK (書籍)，OW (白書)，OM (国会議事録)，OC (Yahoo!知恵袋)の 4 つのサブコーパスについて，それぞれの頻度情報・出現割合の情報を出力できる(どの数値を出力するかはオプションで指定することができる)。その結果，分析対象のコーパスについて，レマの頻度，2 グラムのレマの頻度，2 グラムのコロケーション情報を BCCWJ の 5 種の集合(コーパス全体または 4 つのサブコーパス)と比較できる。

３．今後の開発・応用の方向性

他のコーパスを比較・評価するための資料としての均衡コーパスの有効性を論じる場合，以下のような基本的な問いに答える必要がある。

- どのようなサイズのコーパスデータでもその語彙的特徴を適切に比較できるか。
- どのような指標がコーパス間の比較に適するか。
- 機能語と内容語のような，出現頻度の大きく異なる語彙に同一の統計指標が適用できるか。
- どのような情報を組み合わせることで最も効果的に語彙情報を読み取ることが可能か。
- どのようなインターフェースを使うことでユーザが語彙分析を手際よく進められるか。
- BCCWJ との比較により得られた語彙・文法特徴をどのように応用できるか。

これらの問いに対する答えは，大小さまざまなコーパスを BCCWJ と比較対照しながら模索していく必要がある。本稿はこれら語彙・文法情報のプロファイリング(profiling)の手法とその活用方法の研究に取り組むための出発点と位置づけることができよう。

る。ActivePerl の場合，PPM (Perl Package Manager)を用いることで SQLite の動作に必要なモジュール(DBI, DBD::SQLite)の導入状況を確認し，必要に応じて簡単に追加・更新することができる。詳細は分析ツールに付属するマニュアルを参照されたい。

なお，本稿では短単位情報のみを扱う語彙・文法情報分析システムの構築を報告したが，BCCWJ はその言語単位として，検索や分析の目的に応じ長単位と短単位を使い分けることを当初から想定しており(伝ほか 2007)，教育等の目的には短単位よりも長単位のほうがふさわしい場合が多い(*cf.* 山内 2009)。現在，長単位の仕様はほぼ固まってきており(小椋ほか 2010[3])，今後長単位情報を付与したコーパスが普及していくものと考えられる。

文献

小椋秀樹ほか (2010[3]),『「現代日本語書き言葉均衡コーパス」形態論情報規程集』(第 3 版，特定領域「日本語コーパス」データ班研究成果報告書 JC-D-09-02).

川村よし子 (2000),「インターネット時代に対応した読解教育」(『新世紀之日語教学研究国際会議論文集』), 東呉大学, 中華民国, pp. 347-365. (http://language.tiu.ac.jp/taiwan.pdf)

後藤斉 (2003),「言語理論と言語資料―コーパスとコーパス以外のデータ―」『日本語学』22/5: 6-15.

近藤明日子 (2008),「特徴度の設定」(特定領域「日本語コーパス」言語政策班中間報告書 JC-P-08-01), pp. 13-16.

齊藤俊雄ほか(編) (2005[2]),『英語コーパス言語学：基礎と実践』(改訂新版), 研究社.

大学英語教育学会基本語改訂委員会(編) (2003)『大学英語教育学会基本語リスト JACET List of 8000 Basic Words』大学英語教育学会.

田中牧郎 (2008),「語彙レベルの設定」(特定領域「日本語コーパス」言語政策班中間報告書 JC-P-08-01), pp. 7-12.

田中牧郎 (2009),「言語政策に役立つ，コーパスを用いた語彙表・漢字表などの作成と活用」『人工知能学会誌』24/5: 665-672.

深田淳 (2007),「日本語用例・コロケーション抽出システム『茶漉』」『日本語科学』22: 161-172.

伝康晴ほか (2007),「コーパス日本語学のための言語資源―形態素解析用電子化辞書の開発とその応用」『日本語科学』22: 101-123.

仁科喜久子 (2000),「オンライン教材『あすなろ』プロジェクト」『東工大留学生センター年報』5: 43-45.

橋本直幸, 山内博之 (2008),「日本語教育のための語彙リストの作成」『日本語学』27/10, 50-58.

前川喜久雄 (2006),「特定領域研究『日本語コーパス』のめざすもの」(特定領域「日本語コーパス」平成 18 年度全体会議予稿集), pp.1-8. (http://www2.ninjal.ac.jp/kikuo/tokutei_H18_1.pdf)

前川喜久雄 (2007),「コーパス日本語学の可能性―大規模均衡コーパスがもたらすもの―」『日本語科学』22: 13-28.

丸山岳彦 (2009),「日本語コーパスの現状」『国文学解釈と鑑賞』74/1: 122-130.

山内博之 (2008),「形態素解析に関する提案―日本語教育の視点から―」(特定領域「日本語コーパス」日本語教育班研究成果報告書 JC-E-07-01), pp. 84-93.

山内博之(編) (2008),『日本語教育スタンダード試案 語彙』ひつじ書房.

Biber, Douglas *et al*. (1998), *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Kilgarriff, Adam (2001), “Comparing corpora,” *International Journal of Corpus Linguistics*. 6/1: 1-37.

関連 URL

「あすなろ」(日本語読解学習支援システム)：http://hinoki.ryu.titech.ac.jp/asunaro/index-j.php

「茶漉」(日本語用例・コロケーション抽出システム)：http://tell.fll.purdue.edu/chakoshi-wiki/

「リーディング チュウ太」(日本語読解学習支援システム)：http://language.tiu.ac.jp/

ActivePerl ダウンロード(ActiveState 社)：http://www.activestate.com/activeperl/downloads

SQLite (関係データベース)：http://www.sqlite.org/

UniDic (形態素解析辞書)：http://www.tokuteicorpus.jp/dist/

本稿で構築した語彙情報分析システムの公開情報

日本語教育班ホームページを参照されたい。URL: http://www.tokuteicorpus.jp/g_teaching/