大規模均衡コーパスを利用した語彙・文法情報の評価とその応用

千葉 庄寿 (麗澤大学)
schiba@reitaku-u.ac.jp

1 語彙・文法情報の評価基盤としてのBCCWJ

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(Biber *et al*.1998)。日本の英語教育においても, 基本語リスト(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大規模コーパスを語彙教育に応用した事例がみられる。

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, 日本語のコーパスとしては初の大規模な「均衡コーパス」である(前川 2007:14; 丸山 2009:129)。定量的な研究に役立つBCCWJの設計思想は，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語彙研究に画期的な活路を開くことが期待できる。

「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) という任務に際し，BCCWJ の応用研究を行う言語政策班は，形態素解析されたデータを用いた語彙の計量，特に特徴語抽出の手法について検討をおこなっている(近藤 2008)。また，BCCWJ に基づく日本語教育のための語彙リストの作成の試みも始まった(橋本ほか 2008; 山内(編) 2008)。

しかし，語彙の計量は語彙表の作成にとどまらない。現代日本語を代表する「書き言葉のサンプル」としてのBCCWJの設計思想は，BCCWJそのものの分析だけでなく，他のコーパスデータを評価する比較・評価のための資料としても力を発揮するはずである。当然，以下のような問題点・疑問点が浮かぶ。

- 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の比較・評価に適するか。
- どのような情報を組み合わせることで最も効率よく語彙情報を読み取ることができるか。どのようなインターフェースがよいか。

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と比較対照する作業なしでは得られない。

本ポスターでは，BCCWJ モニター公開データ(2009 年版)を短単位辞書 UniDic [1] (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の語彙・文法情報の比較を手軽に行うシステムの構築を報告する。日本語教育における教材の開発と評価への活用を事例としてとりあげ紹介するとともに，より広い応用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な情報を今後どのように活用できるかを模索する。

オンラインで日本語教材に語彙情報を付与する試みには日本語読解学習支援システム「リーディング チュウ太」[2] や多言語対応日本語読解学習支援システム「あすなろ」[3] などがある。しかし，いずれもBCCWJの語彙・文法情報を利用してデータの分析をおこなうものではない。

また，コロケーション情報の検索を含む語彙分析のオンラインツールとして「茶漉」[4] (深田 2007)があるが，自前のデータを解析する目的には利用できない。

[1] http://www.tokuteicorpus.jp/dist/
[2] http://language.tiu.ac.jp/
[3] http://hinoki.ryu.titech.ac.jp/asunaro/index-j.php
[4] http://tell.fll.purdue.edu/chakoshi-wiki/

2 BCCWJ 語彙情報データベース

BCCWJ の語彙情報データベースは，扱いが簡単な関係データベースエンジンである SQLite 3.7 で構築し，Perl (CGI)，PHP (Web サービス)，.NET Framework (スタンドアロン)により目的に合わせツールを構築している。

現時点で実装している機能は以下の 3 種類である。

- レマ lemma の頻度 : 短単位の語彙素と品詞のペアをキーとして BCCWJ の頻度を検索し，数値を LLR (対数尤度比, Log-Likelihood Ratio, *cf*. Kligarriff 2001; 近藤 2008)で比較する。
- 2 グラム bigram 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。
- 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は収録しておらず，語彙素情報と品詞情報のみを扱っている。

これらの情報に加え，各計量ツールは分析対象のコーパスデータの出現文書数をもとに，各語彙情報の出現割合を出力する。これにより，複数の文書からなるデータをまとめて分析している場合，例えば，政治・経済用語の偏りなど，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。

分析にあたっては，分析対象のデータを 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」を使って解析し，結果を事前にファイルに出力しておく必要がある。BCCWJ の語彙情報データベースと分析データの解析に全く同じ解析環境を使うことにより，出力結果を齟齬なく評価することができるわけである。

なお，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設計の時点で想定しており，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)，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。

参考文献

小椋秀樹ほか (2010)『「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』データ班研究成果報告書 (JC-D-09-02).

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」言語政策班中間報告書 (JC-P-08-01). Pp. 13-16.

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」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

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 (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―」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.

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