言語研究のインフラ整備
—日本語コーパスから見えてきたもの—

前川　喜久雄（言語資源研究系）

0. はじめに

現実に用いられている言語の仕組みをデータに基づいて調べることは容易ではありません。対象が膨大すぎて個人では調査しきれないことが多いからです。この問題を多少とも解消するために国立国語研究所は従来から一連のコーパスを開発してきました。この6年間の成果としては、日本語初の均衡コーパスである『現代日本語書き言葉均衡コーパス』を公開して現代語研究に新しい流れを確立したことのほかに、ネット上の日本語を母集団とする『国語研ウェブコーパス』の構築を完了したこと（公開は来年度）と過去の日本語を対象とする『日本語歴史コーパス』の公開に着手して一部を公開したことが挙げられます。

1.『現代日本語書き言葉均衡コーパス』(BCCWJ)

現代日本語の全体像を把握するために構築した我が国初の均衡コーパスで、規模は1億語です。2010年の公開後、国内外で3,000名以上の研究者によって利用されています。年間100万回ほど検索されており、論文での引用も通算600件を超えています。本コーパスの特徴は、書籍、雑誌、新聞、白書、教科書、ブログ、ネット掲示板、法律、韻文など、現代日本語の書き言葉のさまざまなレジスターを幅広く収録していることにあります。サンプルはすべて形態素解析されており、形態論情報を利用した検索が可能です。語の単位としては、国語辞典の見出し語に該当する短単位と複合語に該当する長単位の二つの単位が利用可能です。

BCCWJを利用することで、従来は誤用ないし例外的と見なされてきた言語事象の生起頻度が意外に高いことが明らかになってきています。

2.『国語研日本語ウェブコーパス』(NWJC)

1億語という規模は利用目的によっては十分ではありません。例えば「書ける」の意味で「書けれる」を用いる例（二重可能あるいはレ足す言葉）はBCCWJには見つかりません。この種の用例は書き言葉には存在しないのでしょうか、それとも生起頻度が低いために1億語ではヒットしないだけでしょうか。この問題を解くにはより大規模のコーパスが必要です。

NWJCはBCCWJの量的不足を補完するために構築した200億語規模のコーパスです。NWJCを検索すると「書けれる」等が100例近く見つかります。ただしNWJCの母集団はウェブ上の日本語テキストなので、BCCWJのような均衡性はありません。またウェブでは同一文が繰り返される（コピーされる）ことが多いので、文単位の単一化（同じ文は1回しか採録しない）を施しています。サンプルは短単位で形態素解析されていますが、さらに文節を単位とした係り受け構造解析が施されており、係り受け構造を検索することが可能です。

3.『日本語歴史コーパス』(CHJ)

国立国語研究所は1948年の創立以来、現代語（とその基盤としての近代語）だけを研究対象としてきましたが、2009年秋の大学共同利用機関法人への移管に際し、歴史的な日本語も研究対象となりました。CHJは奈良時代から現代に到る日本語の歴史的変遷を明らかにすることを目的としたコー

パスで、現在は文学作品を対象に構築を進めています。

2014年には「平安時代編」(源氏物語、古今和歌集、更級日記など14作品、73.8万語)を、2015年には「室町時代編」の一部(虎明本狂言集23.5万語)を公開しました。また近代語についても、従来の『太陽コーパス』に加えて「明六雑誌コーパス」(18万語)と「国民之友コーパス」(101万語)を新規公開しました。これらにはBCCWJと互換性を保った形態論情報が付与されていますので、現代語との比較が可能です。

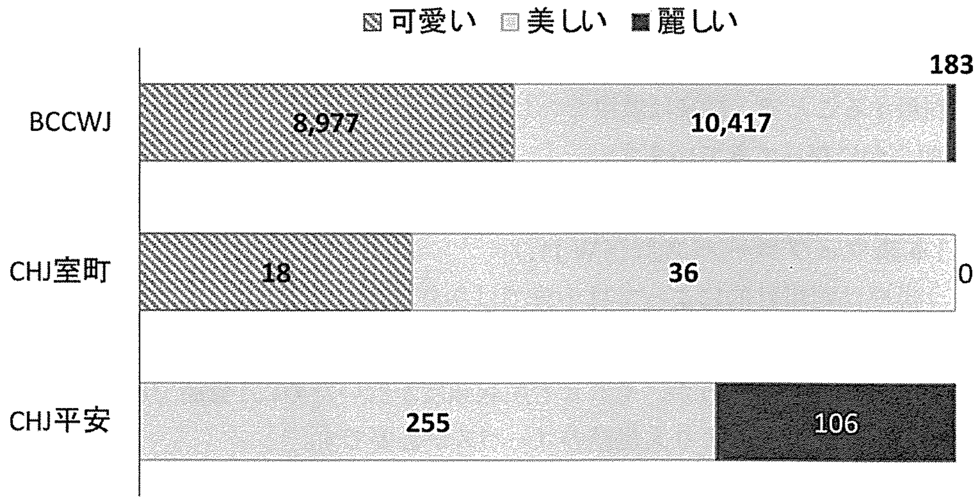
4. BCCWJとNWJCの検索例

非正用とみなされる事象についてBCCWJとNWJCの検索結果をいくつか示します。

項目	BCCWJ	NWJC
形容詞+デス	11,000	393,619
動詞+デス	232	7,172
～シナ(サ)ソウ	75	312
～シナイベキ	11	205
～サセラレ／ラレサセ	479/0	7450/6

5. CHJとBCCWJの検索例

CHJとBCCWJを利用することで歴史的な意味変化の量的側面を確認した例を示します。



参考文献：

前川(監修)『講座日本語コーパス』(全8巻，既刊5巻) 朝倉書店，2013-2016

M. Asahara et al. "Archiving and Analyzing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan". *Alexandria*, 25 (1/2), pp.129-148, 2014

近藤・田中・小木曽(編)『コーパスと日本語史研究』ひつじ書房，2015