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(2007年10月) 13-28

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——

前川　喜久雄
(国立国語研究所)

キーワード
現代日本語書き言葉均衡コーパス，代表性，類義語，コロケーション，文法性判断

要　旨
本稿の前半では筆者らが現在構築を進めている『現代日本語書き言葉均衡コーパス』(BCCWJ)の概要と特徴を紹介し，後半ではBCCWJやそれをさらに発展させた大規模均衡コーパスが言語研究にどのような影響を及ぼすかについての予測を述べた。類義語の研究やコロケーションの研究のように，従来から行われてきた研究がコーパスによって一層進展すると期待されるcorpus-basedな研究のほかに，コーパスなくしては行いえないcorpus-drivenな研究も考えられる。その一例として，文法性判断の個人ないし状況による異同について考察し，文法性判断は言語刺激との接触経験によって容易に影響を蒙ることを指摘した。最後に文法性判断の異同をコーパスによって説明するためには最低でも数十億語規模のコーパスが必要になることを指摘した。

1. はじめに

コーパスを用いた言語研究が世界的に隆盛を迎えている。短く見て1980年代後半以来，長めに見れば1960年代以降のトレンドである。我が国では近年，英語学の領域でコーパスを利用した研究が活発におこなわれている。しかし肝心の日本語を対象としたコーパス言語学的研究はどうかと言うと，お世辞にも盛んとは言いがたい。

言語研究の方法としてのコーパス言語学の特徴としては，実際に用いられた用例を重視することと，定量的な分析を重視することが指摘されることが多いようである（Kennedy 1998；McEnery, Xiao, and Tono 2006)。私は，これらに加えて研究の再現性が確保されうることも指摘しておきたい（前川 2002)。公開されたコーパスを利用した研究の成果は，第三者がそれを再現したり，異なった条件で再分析してみたりすることが可能になるということである。

これらはいずれも伝統的な日本語研究（国語学）において，領域により濃淡はあるにせよ，重視されてきたことがらであるから，日本語研究とコーパス言語学の相性が悪いとは思えない。合理主義的な生成文法理論にくらべれば，経験主義的なコーパス言語学は国語学者にとってはるかに受け入れやすいと思われる。それにもかかわらず日本語のコーパス言語学的研究が不活発である根本原因はコーパスの整備不足にもとめるべきだろう。

筆者の専門は音声学だが1999年以来国語研究所における職務として日本語コーパスの開発に携

わるようになった。1999年から2003年までは科学技術振興調整費の支援をうけて，情報通信研究機構（当時は総務省通信総合研究所），東京工業大学と共同で『日本語話し言葉コーパス』（CSJ: Corpus of Spontaneous Japanese）を開発した。CSJは音声認識研究への応用を念頭において開発したコーパスであるが，言語研究にも利用できるようにさまざまな工夫がこらされている（前川 2004）。

その後2006年度からは研究所にコーパス開発を任務とする言語資源グループが発足することとなり，リーダーを仰せつかることになった。同年半ばには文科省科学研究費特定領域研究「日本語コーパス」が採択され，その代表者も務めている。両プロジェクトの共通目的は，1億語規模の『現代日本語書き言葉均衡コーパス』（BCCWJ; Balanced Corpus of Contemporary Written Japanese）を構築し，2011年に公開することである。

本稿前半では現在構築中のBCCWJの特徴を紹介し，後半ではBCCWJやそれをさらに発展させたコーパスが言語研究にどのような新生面を開くかについて筆者の予想を述べることにする。

2. BCCWJ

BCCWJを紹介するにあたって，最初に術語を説明しておこう。コーパスとは体系的に収集されコンピュータ上に蓄積された検索可能な電子化言語資料のことである。コーパスにも様々なものが考えられるが，幅広いジャンルやレジスターのサンプルを格納することを目標としたコーパスを均衡コーパス(balanced corpus)という。均衡コーパスに格納するサンプルを選ぶ方法も様々であるが，対象となる母集団を明確に規定して確率論的なサンプリングを実施することができれば，対象言語の統計的特性をゆがみなく反映したコーパスを構築できる。このような均衡コーパスには統計的な代表性(statistical representativeness)が認められると言う。私たちはBCCWJを少なくとも部分的には統計的な代表性を備えた均衡コーパスとして設計した。

図1に示すようにBCCWJは3種のサブコーパス(SC)から構成されている。そのうち「出版（生産実態）SC」と「図書館（流通実態）SC」は統計的母集団から無作為抽出されたサンプルからなるコーパスである。出版SCの母集団は2001年から2005年の間に出版された書籍，雑誌，新聞のうち我々が妥当と認めた出版統計に記載されている出版物の全体である。それらの書籍，雑誌，新聞の総文字数を推定し，すべての字が等しい確率で抽出されるような条件のもとで総計約3500万語分のサンプルを抽出した。詳細は丸山・秋元(2007)参照。

図書館SCの母集団は東京都下52自治体の公立図書館に収蔵されている書籍で，1986年から2005年の間に出版されており，ISBNを付与されているもののうち，13自治体以上の図書館が所蔵している書籍の集合である。約34万冊がこれに該当する。ここでも総文字数を推定したうえで無作為抽出を実施して約3000万語分のサンプルを抽出する予定である。

母集団からサンプルを抽出するにあたっては，サンプル長（文字数）を決める必要がある。BCCWJでは固定長と可変長の二種類のサンプルを作成する。前者は1000字固定のサンプルであり，計量語彙論の研究に利用することを想定している。語の頻度，語種の比率，漢字の使用率などである。1000字固定という性格上，サンプルの終端が文などの言語学上の単位と一致するとは

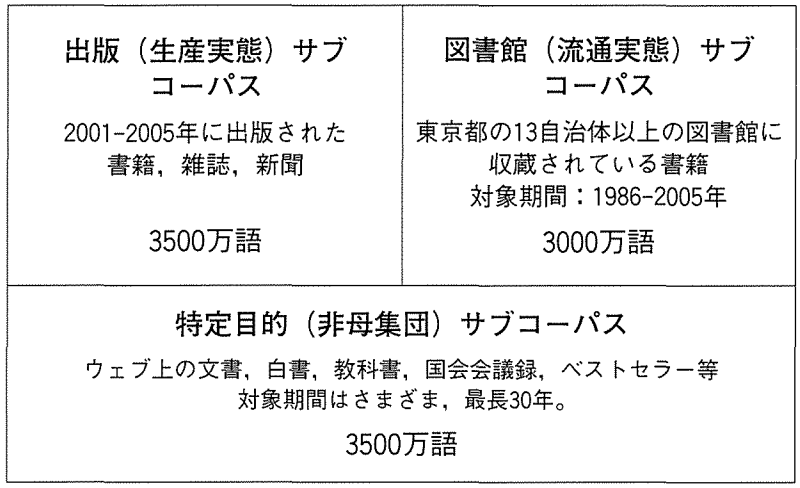

図1　『現代日本語書き言葉均衡コーパス』の構成

限らない（始端は必ず文頭を選ぶ）。

一方，可変長サンプルは「節」や「章」などの文章構成上の単位をサンプルとするものであり，サンプルは始端も終端も言語学上の単位と一致する。一般の言語学的研究にはこのサンプルが利用されることになるだろう。ただし文章構成上の区分が明示されていない文章の場合，作品全体がサンプルとなることを避けるために，サンプル長は10000字を上限とすることにしている。可変長サンプルの平均長は日本十進分類9番台（文芸）の書籍では約4500字である。

最後に図1下部の「特定目的（非母集団）SC」は，上述の二種類の母集団には含まれないか，頻度が低いために無作為抽出による収集が実際上不可能だが言語研究上の必要性が高い言語資料，あるいは国語研が近い将来に実施を予定している言語政策関連の研究課題のために必要とする言語資料のコーパスである。インターネット上の掲示板のテキスト，国会の会議録のような話し言葉の転記テキスト，白書，検定教科書，日本語教育教科書などを格納する予定である。

サンプルの電子化にあたっては，文字セットとしてJIS2008（いわゆるJIS第1第2水準）にJIS2013のサブセット（主に漢字部分）を追加したものを採用し，UTF16でコード化する予定である。BCCWJの設計に関しては山崎(2007)参照[1]。

BCCWJは公開を前提としたコーパスであるから，すべてのサンプルに著作権処理を施す。BCCWJの構築においてもっとも予断を許さないのが，この著作権処理作業である。個人情報保護法の施行によって，2005年以降は著作権者の連絡先を知ること自体が極度に難しくなっているのがその原因である[2]。

2006年度末までに著作権処理を終了したデータには，政府刊行白書から無作為抽出した500万語分のテキスト，ヤフー株式会社提供の「Yahoo! 知恵袋」，そして，国会図書館提供の「国会会議録」（過去30年分）などがある。「Yahoo! 知恵袋」と「国会会議録」にはいずれも単体で1億語を超える量のデータが記録されていると予想されるが，BCCWJではそこから無作為抽出した

一部のデータだけを特定目的サブコーパスに格納している。

著作権処理を円滑に進めるためには，著作権者から提供してもらうテキストが実際にどのように利用されるかを理解してもらうことが肝要である。そのために著作権処理が終了したデータの一部をインターネット上で試験公開している。本稿執筆の時点では，上述の白書500万語と「Yahoo! 知恵袋」500万語の合計1000万語分のテキストに対して簡単な全文検索を試すことができる[3]。

3. 均衡コーパスへの批判

BCCWJのようなサンプルコーパス（特定時点の母集団に依拠した静的なコーパス）を構築するのは時代遅れだという批判を耳にすることがある。批判の論拠となるのは新聞記事データベースの存在とウェブ上のテキストの存在である。

従来，我が国のコーパス言語学では均衡コーパスに代えて新聞記事データベースを利用することが多かった。1994年に自然言語処理研究者の要請に応じて日経新聞と毎日新聞が記事データを研究用に公開したことに端を発し，現在では日経，毎日，朝日，読売，各社の記事データが有償で公開されており，その総量はおそらく10億語を突破している。これはたしかにBCCWJなどよりも大きなコーパスである。

しかし，だから均衡コーパスは必要ない，ということにはならない。新聞記事の日本語には，語彙・文体・表記の面で新聞社特有の統制がくわえられているからである。いわゆる差別語はもちろん，下品な語や過激な語，革新的な文法特徴（例えばラ抜き言葉）は用いられないし，送り仮名などの表記上のゆれはきわめて少ない。漢字も基本的には常用漢字の範囲に収まっており，そこから逸脱する場合も各社校閲部の指針に従っている。全体として，通常の書き言葉に観察される言語変異の大部分が抑圧された日本語となっており，言語のそういう側面を研究するにあたっては不都合が多い。

新聞記事データベースと並んで最近よく利用されるのがウェブの検索結果である。そのため最近では，インターネットという巨大なコーパスが存在し，成長し続けている以上，コーパスとしてはそれを利用すれば済むという批判を耳にするようになった。

WWW上にはたしかに膨大多様なテキストが蓄積されておりGoogleをはじめとする検索エンジンによってそれをすばやく検索できる。文法上何らかの疑義を生じたとき，とりあえずインターネットを検索してみるというのは現代のひとつの研究スタイルであり，私もその恩恵にあずかっている一人である（4.2節のコロケーションについての議論参照）。しかし検索エンジンがもたらすサンプルには以下のような問題がある。これらの問題はコーパス言語学の重要な応用領域である文体論研究や言語変異研究に関しては非常に深刻である。

(1) **検索の再現性の欠如**。WWWは日々変化しつづけているので検索結果も刻々変化する。さらに，それよりも深刻な問題として，検索結果の不安定性がある。検索エンジンによっては同一条件でおこなった検索の結果が短期間に大幅に変化することがある[4]。

(2) **ジャンル・著者情報の欠如**。所与の文書の出典情報，特にジャンル情報（新聞記事か小説

か論文か，論文ならば領域は何か）をあらかじめ知ることができない。また文書の著者の社会的属性（性別，年齢など）を知ることができない。そのためサンプルの言語学的な偏りを評価できない。

(3) **整列条件が非公開。**検索結果がどのような基準で整列(sort)されているかが公開されていない。多数の文書がヒットしたときに上位だけを検討対象にすると偏りが生じる可能性がある。

昨今，日本語自然言語処理で大きな話題となった「Webから自動構築した大規模格フレーム」(河原・黒橋 2006) は検索エンジンには依存していないが，(2)の問題からは免れていない[5]。言語研究のためにはやはりBCCWJのようなサンプルコーパスの構築が必要である。

4. コーパスがひらく可能性

さてそれではBCCWJのような均衡コーパスは言語研究にどのように貢献するだろうか。英語の場合，BNC(*British National Corpus*)のような均衡コーパスがもっとも大きく貢献したのは第2言語としての英語(ESL)教育であったように思える。*COBUILD* や *Longman Dictionary of Contemporary English* に代表される学習者用英語辞書の開発，話し言葉に配慮した文法書(Biber et al. 1999)，コロケーションの分析と教授法(Nesselhauf 2004)などはコーパスなくしては実施できなかった研究である。

先にも指摘した，実例を重視する，対象を定量的に把握するというコーパス言語学の特徴は，外国語教育との親和性が高い。BCCWJの場合も語学教育はもっとも期待される応用領域であり，非母語話者に対する日本語教育とならんで母語話者に対する国語教育への応用も期待されている[6]。以下では狭い意味での言語研究に対する貢献に的を絞って私見を述べることにする。

4.1. 類義語

最近の国語辞書は類義語の記述に力を注ぐようになってきたが，まだ十分とはいえない。大規模コーパスのもたらす豊富な用例とジャンル情報は類義語記述の大きな武器になると思われる。

例として「光景」と「風景」の関係をとりあげる。この2語の異同については国広(1997)の精緻な意味分析があるが，ここでは形態論的な特徴に注目することとする。コーパスを検索すると両者は語形成上のふるまいが著しく相違していることがすぐにわかる。2003年の毎日新聞記事1年分を検索した結果を表1にまとめた。

表1 「風景」と「光景」の複合語になりやすさ

語	総生起数	複合語後部要素としての生起数 および総生起数に対する比率	複合語の異なり語数
風　景	954	259 (27.1%)	107
光　景	514	4 (0.8%)	4

「風景」は954回，「光景」は514回生じているが，「風景」のうち259回は「原風景」「心象風景」「田園風景」「日常風景」のように複合語の後部要素として出現しており，異なり語として107種の複合語が生じている。一方「光景」が複合語後部要素となっていたのは「野積み光景」「日常的光景」「歴史的光景」「神話的光景」の4例のみである。しかもそのうち3例には接尾辞「的」が用いられており，複合語としての熟合度が低いことを示唆している（ちなみに「風景」における「的」は「社会的風景」「歴史的風景」「幻想的風景」の3例であった)。また「風景画」「風景写真」など「風景」を前部要素とする複合語は67例（異なりで11語）生じていたが，「光景」を前部要素とする複合語は皆無であった。以上を要するに「風景」と「光景」とでは複合語になりやすさに顕著な差が認められることがわかった。

これと同じ特徴を備える語の対は少なからず見つかる。表2には「兵器」と「武器」の例を示した。「風景」「光景」と同一条件での検索結果を示すと，「兵器」は7470回，「武器」は1695回生じているが，「兵器」のうち6870回は「核兵器」「大量破壊兵器」「化学兵器」のように複合後の後部要素として出現しており，異なり語として96語の複合語が生じている。一方「武器」が複合語後部要素となっていたのは79回で，異なり語は「小型武器」「携行武器」「国産武器」など21種である。

このような造語法にかかわる情報は日本語学習者にとっては非常に貴重な情報である。もちろん，母語話者対象の国語辞典に記載しても歓迎されるに違いない。この種の情報をコーパスから抽出するのに技術上の困難はないので，大規模コーパスが普及すれば，この種の情報を組織的に記載した辞書もおいおい編纂されるようになるだろう。

表2　「兵器」と「武器」の複合語になりやすさ

語	総生起数	複合語後部要素としての生起数 および総生起数に対する比率	複合語の異なり語数
兵　器	7470	6870（92.0%）	96
武　器	1695	79（ 4.7%）	21

4.2. コロケーション（句の意味の研究）

次に動詞「起きる」「起こる」「生じる」の異同を考えてみよう。やはり意味論的な考察はおこなわないこととして，コーパスから作成したコンコーダンスを検討してみると，主格補語（ガ格補語）にたつ名詞によって動詞の生起率にくっきりとした違いがあらわれる。表3に前節と同じ毎日新聞記事にくわえて国立国会図書館が公開している「国会会議録」の全体を検索した結果を示す。いずれの場合も，「事件が起きた」「問題が生じれば」のようにガ格補語と動詞とが直接隣接しているケースだけを検索した結果である。「問題」はいずれの動詞とも共起するが，「事件」は「起きる」か「起こる」かであって，「生じる」ことは稀である。

表3　「問題」「事件」と「起きる」「起こる」「生じる」の共起関係

コーパス	ガ格補語	起きる	起こる	生じる
新聞記事	問題	84（52.9%）	12（ 7.5%）	63（39.6%）
	事件	301（87.2%）	47（12.5%）	1（ 0.3%）
国会会議録	問題	85（20.7%）	143（34.9%）	182（44.4%）
	事件	100（43.1%）	125（53.9%）	7（ 3.0%）

このような語と語の共起関係における選好性はコロケーション(collocation)と呼ばれる。先にコーパス言語学の方法上の特徴として，実例の重視と量的側面の重視を指摘したが，これらの特徴に照らしても，Firth(1957)が“actual words in habitual company”と説明したコロケーションの研究は最もコーパス言語学らしい研究テーマのひとつだと言える。

現代の文法理論の多くは文の意味が語彙項目固有の意味と統語構造とによって決定されると考えているのだと思うが，コロケーションには，そのような取り扱いを拒む面がある。もうひとつ例をくわえて考えてみよう。

「良い」と「優れた」はいずれも対象が優越した状態にあることを示す連体修飾語として類語関係にある。今「～が｛優れた／良い｝人」ないし「～の｛優れた／良い｝人」というフレーム中の名詞「～」として「性格」「頭」など7語をとりあげて，Googleでのインターネット検索の結果を示すと表4のようである。「性格」と「頭」には「優れた」よりも「良い」が共起しやすいが，「能力」「才能」はその逆のパタンである。さらに「成績」「容姿」のように両者ともかなり用いられているケース，「頭脳」のように両者ともほとんど用いられないケースがある[7]。

表4　「優れた」「良い」といくつかの名詞の共起関係

名詞	～が優れた（すぐれた）人	～の優れた（すぐれた）人	～が良い（よい）人	～の良い（よい）人
性格	1	0	32360	2504
頭	0	9	70610	128700
能力	8786	9690	3	3
才能	7	2246	0	1
成績	4015	3207	12200	646
容姿	399	731	3848	2467
頭脳	5	5	2	9

ここまでに挙げた例に関して大切なのは，「事件が生じる」「性格が優れた人」「頭の優れた人」等の句が誤用であるとは言えないことである。これらの句は生起頻度が低いだけでなく，「事件

が起きる」「頭の良い人」などと比較すれば相対的に不自然と判断されるが，意味が通じないわけではなく，実際に日本語母語話者らしき人によって用いられている[8]。そのような用例が存在する以上，上記の制約を語の選択制限として記述することは望ましくない。

また，これを語用論の問題とみることにも問題がある。語の意味が周囲の環境から影響を蒙っているという点ではたしかに語用論上の問題と言えるのだが，環境と言っても言語的文脈や言語外的発話状況は含まれておらず，単純にふたつの語と語が共起していることだけが環境の本質であると思われるからである。

コロケーションの意味は構成要素となる語の意味の単純結合では予測できない（不透明性がある）点では熟語（慣用表現）に似ているが，完全に固定された表現ではない点で異なっている。「顎を出す」「水をさす」などの熟語は「出した顎」「さした水」に変換すると熟語としての意味が失われてしまう。一方，コロケーションならば「良い性格の人」「性格の良い人」のようにどちらも自然であることが多い。

以上を要するにコロケーションは特定の語と語の間に成立する慣用的な（すなわち文法的に固定化されていない）結合関係であり，その意味には語彙的意味の交互作用(interaction)によって生じる部分的な不透明性が認められる。そのためコロケーションの記述は文でも語でもなく句のレベルで実施するのが妥当である[9]。大規模なコーパスを活用できればコロケーションの候補は或る程度まで自動的に検出できるので，日本語についても今後コロケーションを重視した意味研究が組織的におこなわれ，その成果は辞書や教材類に次第に反映されていくだろう。

4.3. 文法性判断

コーパス言語学の可能性を論じてTognini-Bonelli(2001)はcorpus-based investigationとcorpus-driven investigationの区別を主張している。前者は従来から言語研究において検討されてきた諸問題をコーパスを利用して解決しようとする研究である。一方後者は，コーパスそのもののなかから従来の言語研究では認識されてこなかった現象を発見し，それを解決しようとする研究である。前者にとってコーパスは研究ツールであるが，後者にとってのコーパスは研究対象そのものである。

ここまでに触れてきた事例はいずれもcorpus-basedな研究の色彩が濃いものであったが，corpus-drivenな研究としては，どのようなものが考えられるだろうか。上述の説明に従えば，corpus-drivenな研究は従来の研究から懸絶した問題を扱わねばならないのだから，それを予見することはなかなか難しい。本節と次節では，この問題を考えるひとつの手掛かりとして，文法性(grammaticality)について考えてみることにする。

文法研究では文の文法性判断が重要だが，その判断に個人による（あるいは状況による）異同が生じることがある。文の適格性の判断に幅がありうるという事実は言語の本質を考察するうえで非常に重要である。例えば以下の文の文法性判断を要求されたとき，これを非文と判断する人は少なくないだろう。

(1) 昨晩，あるいは昨夜おそく，このあたりは雨が降ったです

しかし，これは実際に用いられた日本語である。しかも40年以上にわたって60刷を重ねてきたロングセラーに見つかる用例である（グロルラー著，阿部主計訳「奇妙な跡」，江戸川乱歩編「世界短編傑作集２」創元推理文庫，初版1961)。翻訳だから日本語がおかしいのだ…というのはこの場合理屈にならない。翻訳者は立派な日本語母語話者だからである。「だったです」ないし「動詞＋タ＋デス」を手許の資料中に検索してみると話し言葉らしい用例が実際にみつかる。(2)-(4)は「文芸春秋」の座談会[10]，(5)もやはり文芸春秋の対談[11]，(6)は「国会会議録」，(7)は『日本語話し言葉コーパス』中の用例である。もちろんインターネット検索でも類例を発見できる。

(2) まさに正岡子規だったですよ
(3) それだもんで参っちゃったですよ
(4) ああ，これは本腰を入れなきゃいかんと思ったですね
(5) 僕はエボシ御前というのは，早く出てきた織田信長の女性版だと思ったですね
(6) 政府は一体具体的に何をやったですか
(7) 初めて海外に行ったですよ

これらの用例が用いられたであろう文脈を想像してみる。すると私などは(1)を非文と断定しにくく感じられてくる。合理化の契機が与えられれば，むしろ適格文に思えてくる。本例の場合であれば(2)-(7)の用例を発見し，その出典を確認することによって「ああ，話し言葉ならたしかにこう言うこともあるな」と思えてくるのである[12]。

もうひとつ例を挙げよう。(8)は作家今東光(1898-1977)が書いた随筆の一節である[13]。

(8) 僕たちは警察に信頼して好いと思う

私は最初この例に触れたとき誤植ではないかと考えた。しかし「青空文庫」を検索してみると「～に信頼する」の例が次々と見つかる。

(9) 生活を維持するに足る詩的天才に信頼したために胃袋の一語を忘れた
(10) 安心して，僕に信頼したらよかろう
(11) あまりに現在の脆弱な文明的設備に信頼し過ぎているような気がする
(12) まつは，善良で私に信頼し，同時に無智だ

これらはいずれも明治生まれの著名な文筆家の日本語である。(9)は芥川龍之介(1892-1927)の「河童」，(10)は夏目漱石(1867-1916)「二百十日」，(11)は寺田寅彦(1878-1935)の「石油ラムプ」，(12)は宮本百合子(1899-1951)「文字のある紙片」である。その時期の日本人にとっては「～に信

頼する」が適格文であったことがわかる。また，日本国憲法前文にも「〜に信頼する」の例がある[14]。

この場合も一度(9)以下の例を体験してしまうと私は現代語としても(8)を非文とする気がなくなってしまう。自分自身が「〜に信頼する」と書くことはないかもしれないが，(8)を適格文として受容することにこだわりがなくなってしまうのである[15]。

このような文法性判断の異同が何故生じるのかは，それ自体が言語の本質にかかわる大切な研究課題である。殊に(1)-(7)のように，自分がそのように行動していながら，意識的な内省判断では排斥されてしまうような用例の存在は大変興味深い。私は言語研究の方法としての内省（意識）の問題点を指摘してコーパスの必要性を論じたことがあるが（前川 2007b），ことは言語研究だけにとどまるものではなく，人間の認知機構全般に及ぶ可能性がある（下條 1996）。

ちなみに「信頼する」の例については，これを言語の通時的変化とみて文法の共時的記述から除外したい人もいるだろう。しかし読書人は上に掲げた程度の過去の言語資料ならば日常いくらでも触れている。またそもそも現実の言語共同体には構成員の年齢差という通時的要素が抜き難く存在しているのであり，例えば私の年齢（1956年生）の人間ならば(8)や(9)の著者と同世代の祖父母と日常的に接触していた人が多いはずである。そしてそれらの人々は「〜に信頼する」と言ったり書いたりしていた可能性が高い。“実際の用例を重視する”コーパス言語学はこのような事実を無視すべきではなく，むしろ研究対象として積極的にとりあげるべきだろう[16]。また，井上(2001)が示唆しているように，現代語におけるヲ格とニ格の交替はかなりの数の動詞に生じており，なかには通時的変化として説明することが適当でないものもあると思われる。

4.4. 文と非文の境界

従来の言語研究，特に生成文法理論では文と非文の境界は明確に（二値的に）定まるものと考えてきた。しかし文法性判断に異同が存在する状態が稀な例外でないとすれば，文と非文の関係を連続的な変化としてとらえることが考えられる。その場合，文法には正解が存在しないこととなり，文の候補として与えられた文字列の文法性の程度を評価することが新しい文法の主要な目的となるだろう。Corpus-driven な言語学がめざすべき目標には，このような文法性の程度を評価する連続量の計算法と，その評価値の高低が何に起因するかを説明するための理論が含まれていてしかるべきである。

第一の目標については，十分に大きな規模のコーパスがあれば，統計的な言語モデル—隠れマルコフモデルによるものなど—が所与の単語列の生起確率を与えてくれるので，その確率を文法性の測度として利用できそうである。ただし，現在のところそのようなモデルは表層的に生じた単語列の確率をコーパスから計算しているだけであることが多く，そこでは語に特性と文ないし節の統語的な特性とが分離されず渾然一体となって計算されてしまっている。語の特性と構文的な特性とを切り分けることができれば（そしてコロケーションの現象を語の交互作用として位置づけることができれば），言語研究にとって真に興味深い知見を得ることができるだろう[17]。

さて，このような議論をすると，非文の生起確率はコーパスから計算できないと主張する人が

でてきそうである。あらかじめ反論しておくと，そのような主張は生成文法の揺籃期はいざしらず，現代では通用しにくい。十分に大きなコーパスのデータを何らかの形でクラス化しておけば，データを補間することによってコーパス中には観察されていないサンプルの生起確率を或る程度の精度で推定できるからである。Pereira(2000)は，そのような方法で生成文法初期の有名な例文 *Colorless green idea sleeps furiously* と *Furiously sleep ideas green colorless* の生起確率を推定すると，両者間に大きな差が認められることを示している。

次に第二の目標について述べる。この目標を達成するためには文法性判断の異同に関与する要因を調査して，その発生原因を解明しなければならない。調査すべき要因のなかにはレジスターの差，文体の差，執筆者の年齢，出身地，そして次節で論じる各種言語刺激との接触頻度などが含まれる。これらの多くは言語共同体の多様性を生み出す要因として従来から指摘されてきたものであり，社会言語学（特に言語変異研究）の領域で，これまでにも多くの研究が積み重ねられている。しかし残念なことに，従来の研究では多くの研究が個別に累積されているだけで，ひとつの大理論に収斂していない。先行研究の成果を発展させて，文法性判断の異同を説明しうる大理論を構築するためには，現在の社会言語学のように，研究対象としてとりあげる言語現象ごとに異なるデータを分析するのではなく，巨大なデータを共有しておいて，そこに含まれる多くの言語現象を多面的に分析したのちに分析結果を総合して，包括的な理論に到達することが望まれる。

結局ここでもまたコーパスが重要な役割を果たすことになるのだが，社会言語学的分析に用いるコーパスにおいては，テキストおよび言語使用者の属性情報が格別に重要であることを改めて指摘しておきたい。既に指摘したようにウェブ検索エンジンや新聞記事のデータはこの目的のためには明らかに不十分である。第二の目標を達成するためには組織的に設計された大規模均衡コーパスが絶対的に必要である。

5. 超巨大コーパス

最後に少し視点を変えて corpus-driven な研究のためのコーパスはどの程度の規模であるべきかという問題を考えてみたい。既に4.3節でみたように，言語の文法性判断は言語資料との接触経験に影響される。言語運用におけるこのような側面は，従来の言語研究では全くと言ってよいほど顧みられていないが，コーパス言語学では重要な課題として把握すべきである。心理学では刺激との接触頻度が刺激の選好に影響することが知られており，単純接触効果(mere exposure effect)と呼ばれているが(Zajong 1968)，言語がその例外であるとは考えられない。個人によって文法性判断に異同が生じる根本的な原因のひとつは，各個人がどのような言語刺激とどの程度接触しているかの差にもとめられるだろう[18]。個々人がそれまでの人生においてどのような言語資料に接触してきたかについての情報があれば，その他の情報とあわせ用いることによって，その人の文法性判断の特性を予測できるかもしれないのである。

そのような情報を求めることは十数年前までは笑うべき妄想であった。しかし現在ではあながちそうともいえなくなってきている。コーパスが十分に巨大化すれば，個人の言語接触歴をシミ

ュレートできると考えられるからである。4.3節でとりあげた「～に信頼する」のような書き言葉中心の表現であれば，年齢，性別，学歴，専門，趣味，職業，読書傾向などの社会的属性から特定の個人が過去に当該言語表現に接触した確率の期待値を計算できる可能性がある[19]。

さて，そのような計算を可能にするコーパスはどの程度の規模になるだろうか。試みに2005年1年間に私自身が読んだ和書の記録をとってみた。大雑把な推定であるが，年間で約2600万文字，1語あたり平均1.7文字と仮定して語数になおせばおよそ1530万語を読んでいた。

この調査では単行本だけを対象としたので新聞・雑誌の記事，メール，ウェブ上の文書，マンガ，論文，事務書類等は除外されている。それらを含めれば1年で2000万語以上の書き言葉に接触しているだろう。そのような接触状態を30年間継続したと仮定すれば，私がこれまでに接触した言語資料の総体は6億語を超える。BNCやBCCWJのような1億語規模のコーパスでは私程度の読書量の人間の経験すらカバーできないのであった。シミュレーションのためには少なくとも数十億語，望ましくは百億語規模の均衡コーパスが必要であると思われる。

百億語とはどのような規模か。BCCWJの図書館SCの母集団が約280億語であるから，百億語規模のコーパスには，書き言葉を対象にする限り従来のコーパスの母集団のほぼ全体が記録されることになる。インターネットを対象にくわえ，50年程度の時間幅をもたせても，対象の1～2割を記録できるだろう。

現代語のコーパスは誕生以来，実はこの規模を目指して着々と進化してきているように思える。Brown Corpus（1965年公開，100万語），BNC（1995年，1億語），Bank of English（2005年，5.25億語）について，コーパスの公開年をX，収録語数の常用対数をYとして片対数グラフにプロットすると，3点は直線の上にきれいに並ぶ。線形回帰によってこのトレンドを外挿すると英語コーパスは2030年頃に百億語を突破することになる。

今は夢と思えるかもしれないが，著作権の問題さえ解決できれば将来のコーパスは実際に百億語に到達するのではなかろうか。コーパスの真価を発揮させるためにその規模が必要とされるのであれば万難を排してそこまで進もうとするのが研究者だからである。そのようなコーパスは言語研究のあり方を根本的に変えずにはおかないだろう。

付記：「Webから自動構築した大規模格フレーム」について

本稿は2007年3月に開催した特定領域研究「日本語コーパス」公開ワークショップでの講演原稿（前川 2007a），ならびに日本言語学会第134回大会シンポジウムでの講演原稿に加筆したものである。本稿をあらかた執筆した後に，3節で言及した「Webから自動構築した大規模格フレーム」がウェブ上でβ公開されていることを知った[20]。格フレーム構築の際に利用された実際の用例が表示可能になっている点が画期的であり，これは一種のコーパスとみなすことができる。格フレーム自動構築のために検索したサンプルは約5億文であるから，数十億語規模のデータが検索されていることは間違いなく，5節に述べた「超巨大コーパス」を近似していると考えられる。

試みに4.3節で取り上げた「信頼する」を検索してみた。文字列「信頼」に関しては動詞述語

として158種類の格フレームが登録されており，そのうち13種類にニ格の補語が含まれていた。そのうち3種はヲ格と交代することのない例（「無条件に信頼出来る」，「初対面の人に信頼して貰うための努力を」「私共は，お客さまに信頼して頂いていると思っております」）であったが，残り10例はヲ格と交代しうる例であった。数十億語規模のコーパスでは，4節でとりあげた類の文法現象も corpus-based に研究できる可能性が高いことがわかる。

ただし問題もあった。上記の10例を精査してみると，1例が日本国憲法前文であった他は，すべて宗教関係（特にキリスト教関係）のテキストであった[21]。4節で取りあげた(8)-(12)のような文芸作品中の用例は全く含まれていない。ウェブの検索方法や格フレームの自動抽出アルゴリズムに起因する問題なのだろうが，ウェブだけを対象として構築したコーパスと BCCWJ のような，より伝統的な方法で構築したコーパスとではサンプルの言語的性格がかなり異なってくることがあるようだ。今後 BCCWJ の構築が進んだ段階で両者をきちんと比較してみる必要がある。

注

1 BCCWJ 関連の論文のうち主要なものは特定領域「日本語コーパス」のウェブサイトからダウンロードすることができる（http://www.tokuteicorpus.jp/）。

2 フェアユースの概念のない日本において著作権者の権利保護と公共の利便を確保するためには著作権者と連絡がとれることが大前提である。個人情報保護法はこの連絡の可能性を大幅に狭めている。

3 http://www.kotonoha.gr.jp/demo/

4 一例として2007年4月中旬と5月中旬に「容姿が優れた人」という句を Google で検索した際のヒット件数を比較すると4月が399件，5月が10件であった。4.2節の議論参照。

5 ウェブ上のテキストの偏りも問題になる。これについては本稿の「付記」参照。

6 特定領域研究「日本語コーパス」では計画班のなかに日本語教育班，言語政策班を設置して，教育領域での応用を試みている。

7 「頭脳」の場合，多いのは「頭脳明晰な人」である（8770例）。

8 「事件が生じる」は『環境白書』や『外交白書』に用例がみつかる。

9 コーパス言語学では phraseology という用語が用いられる。Mezzo-structure の意味といってもよい。

10 文芸春秋1989年2月号「吉例新春句会」。

11 文芸春秋1997年11月号「『もののけ姫』ラストシーンの謎を解く」。

12 用例のジャンルが重要な所以である。3節の議論参照。

13 今東光「赤線消ゆ・東光辻説法」半藤利一編『「文芸春秋」にみる昭和史（三）』文芸春秋，1988（初出 1948）。

14 「日本国民は，（中略），平和を愛する諸国民の公正と信義に信頼して，われらの安全と生存を保持しようと決意した。」小木曽智信氏の教示による。

15 著名な作家による例を示されると文法判断への影響が強いことは認知心理学の「認知的不協和理論」によって説明できそうである（下條 1996参照）。

16 柏野(2006)は国語辞典の釈義および用例との関係でこの問題を論じている。

17 Fujimura(1968)はこのような考えを非常に早い時期に表明している。

18 横山(2006)は異体字の選好を単純接触効果の観点から研究している。そこに示された「文字生活の俯瞰図」(p.200)は，文法性の問題にも示唆するところがある。

19 このような研究で本当に重要なのは話し言葉だろうが，話し言葉について数十億語規模のコーパスを構築することは現在でもまだ見果てぬ夢に属する。ただし録音するだけならば，個人が一生涯に接する程度の音声は圧縮してハードディスクに保存可能である。やがては音声認識技術の発展によって，保存された音声を実用的な精度で自動認識することも可能になるだろう。

20 http://nlp.kuee.kyoto-u.ac.jp/nl-resource/caseframe.html

21 「彼が常によく祈ったというのは，彼が全く神に信頼していたからです。」「信仰があるというのは神の力にすべてを信頼することなのです。」「そしてそのイエス様の言葉に信頼することが，信仰の出発点なんです。」等々。

参考文献

井上優(2001)「問13」『新「ことば」シリーズ14言葉に関する問答集—よくある「ことば」の質問—』，36-37，国立国語研究所

柏野和佳子(2006)「国語辞典の釈義と用例の検討」『言語処理学会第12回年次大会予稿集』，S1-2，言語処理学会

河原大輔・黒橋禎夫(2006)「高性能計算環境を用いたWebからの大規模格フレーム構築」『情報処理学会 自然言語処理研究会』171(12)，67-73，情報処理学会

国広哲弥(1997)『理想の国語辞典』大修館書店

下條伸輔(1996)『サブリミナル・マインド—潜在的人間観のゆくえ』中央公論社

前川喜久雄(2002)「『日本語話し言葉コーパス』を用いた言語変異研究」『音声研究』6(3)，48-59，日本音声学会

前川喜久雄(2004)「『日本語話し言葉コーパス』の概要」『日本語科学』15，111-133，国立国語研究所

前川喜久雄(2007a)「特定領域『日本語コーパス』—目標，進捗状況，そして夢—」『特定領域研究「日本語コーパス」平成18年度公開ワークショップ(研究成果報告会)予稿集』1-13

前川喜久雄(2007b)「内省からコーパスへ」『文部科学教育通信』169，22-23，ジアース教育新社

丸山岳彦・秋元祐哉(2007)「『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法—現代日本語書き言葉の文字数調査—」『国立国語研究所内部報告書(LR-CCG-06-02)』国立国語研究所

山崎誠(2007)「『現代日本語書き言葉均衡コーパス』の基本設計について」『特定領域研究「日本語コーパス」平成18年度公開ワークショップ(研究成果報告会)予稿集』，127-136

横山詔一(2006)「異体字選好における単純接触効果と一般対応法則の関係」『計量国語学』25(5)，199-214，計量国語学会

Biber, D., S. Johansson, G. N. Leech, S. Conrad, & E. Finegan(1999) *Longman grammar of spoken and written English*, London: Longman.

Firth, J. R.(1957) A synopsis of linguistic theory 1930-1955, *Studies in linguistic analysis*, 1-32, Oxford: Blackwell (Reprinted in F. R. Palmer(ed.)(1968) *Selected papers of J. R. Firth*, 1952-1959, Harlow: Longmans).

Fujimura, O.(1968) Approaches toward a model of linguistic behavior, *Annual Bulletin Research*

Institute of Logopedics and Phoniatrics 7, 42-45, University of Tokyo.

Kennedy, G.(1998) *An Introduction to corpus linguistics*, London: Longman.

McEnery, T., R. Xiao, & Y. Tono(2006) *Corpus-based language studies: An advanced resource book*, London: Routledge.

Nesselhauf, N.(2004) *Collocations in a learner corpus (Studies in Corpus Linguistics: 14)*, Amsterdam /Philadelphia: John Benjamins.

Pereira, F.(2000) Formal grammar and information theory: Together again?, *Philosophical Transactions of the Royal Society* 358(1769), 1239-1253.

Tognini-Bonelli, E.(2001) *Corpus linguistics at work (Studies in Corpus Linguistics: 6)*, Amsterdam/ Philadelphia: John Benjamins.

Zajong, R. B.(1968) Attitudinal effects of mere exposure, *Journal of Social Psychology, Monograph Supplement* 9, 1-27 (Reprinted in R. B. Zajonc(ed.)(2003) *The selected works of R. B. Zajonc*, Wiley & Sons).

(投稿受理日：2007年6月6日)

前川　喜久雄（まえかわ　きくお）
国立国語研究所研究開発部門
190-8561　東京都立川市緑町10-2
kikuo@kokken.go.jp

Prospects of Japanese corpus linguistics:

The influence of large-scale balanced corpus

MAEKAWA Kikuo
The National Institute for Japanese Language

Keywords
Balanced Corpus of Contemporary Written Japanese, representativeness
synonym, collocation, grammaticality judgment

Abstract

The aim of this paper is twofold. In the first half of the paper, design issues of the *Balanced Corpus of Contemporary Written Japanese* were discussed paying special attention to the recent criticisms against the manual (as opposed to 'automatic') construction of a large-scale balanced corpus. The last half of the paper is devoted to the discussion about the influence of a large-scale balanced corpus on the linguistic study of the Japanese language, encompassing both traditional (corpus-based) and innovative (corpus-driven) research themes. Analyses of synonyms and collocations were presented as the examples of traditional research topics. Also, investigation about the inter-personal and/or situational differences of grammaticality judgment was presented as an example of innovative research topics. Lastly, issues about the corpus size were discussed from a point of view of the coverage by corpus of the total reading experience of a particular person. It turned out that the size of corpus should be considerably larger than one billion words if we want to construct a theory about the inter-personal difference of grammaticality judgments.