Corpus linguistics and Japanese language studies

コーパス言語学と日本語研究

後藤　斉
(東北大学)

キーワード
コーパス，コーパス言語学，現代日本語書き言葉均衡コーパス

要　旨

本稿は，コーパス言語学をもっとも発達させたイギリスにおける事情と日本におけるコーパス研究の位置づけとを対比しつつ歴史的に概観して，その発展の違いの要因を探り，あわせて今後に対するなにがしかの見通しを得ようとするものである。イギリスにおいてコーパス言語学が発達したことには，主要因としては言語研究の流れに沿うものであったことが挙げられ，ほかにもいくつかの言語内的および言語外的要因が挙げられる。それに対して，日本では，計算機利用の言語研究の歴史は長いが，コーパスの概念の精緻化には至らず，現在，代表性を備えていて，人文系の研究者が共有できるようなコーパスが存在しない。現在の不十分なコーパスでも意味論の研究などに利用することが可能ではあるが，国立国語研究所が「現代日本語書き言葉均衡コーパス」の構築に着手したことの意義は大きい。ただし，それを十分に生かすためには，利用者の側にも主体的な努力が求められる。

1. はじめに

本稿は，コーパス言語学をもっとも発達させたイギリスにおける事情と日本におけるコーパス研究の位置づけとを対比しつつ歴史的に概観し，あわせて，今ようやく本格的な姿を取り始めている日本語コーパス言語学の基礎をより強固なものにするために，今後に対するなにがしかの見通しを得ようとするものである。筆者は後藤(1995)において日本語コーパス言語学に対するその時点での見通しを述べたが，この十数年においてそれが順調に実現されたとは言いがたい現状にある。本稿では，この点について改めて考察してみたい。

2. Leech(1984)をめぐって

1984年10月13日，東北大学で開かれた第89回日本言語学会大会において，当時神戸大学に客員教授として滞在していたランカスター大学のG.リーチ教授は，“The value of a corpus in linguistic research: a reappraisal”(「言語学研究におけるコーパスの価値：再評価」)と題する公開講演(Leech 1984)を行った。リーチは意味論の研究者として日本においてもすでに著名であったが，この講演では，コーパスの価値に関するチョムスキーの側からの否定的な見解に対して理論的な観点から反論した上で，類義語の対である副詞almostとnearlyおよび動詞seemとappearを例として取り上げ，ランカスター大学などで開発されたLOBコーパスから引かれた多

数の例文を挙げることによって，コーパスの有用性を具体的に示した。

リーチは，コーパスのデータから示される，それぞれの語が生起しやすいテキストのジャンルに違いがあることやそれぞれの語と共起する語の意味的な範囲に違いがあることなどに注目し，ここからそれぞれの類義語の対において almost と seem は無標であり一方 nearly と appear は有標であるという形で両者の関係をまとめる。ところで，このような違いは，リーチのような意味論のすぐれた研究者にとってさえ，ネイティブ・スピーカーとしての直感のみからは容易に得られるものではない。このことを基にして，リーチは言語記述におけるコーパスの価値を再評価すべきであると論じたのであった。同時に，リーチは慎重にも，コーパスのデータの分析にあたっては内省も必要であることを付け加えている。

この講演は，コーパスの中から特定の語を検索した上でそれが生起する環境に注目して，それぞれの語彙項目の性質を探るというコーパス言語学の基本的な手法とその理論的な基盤を論じたものとして，現在でもその根本における意義を失っていないと思われる。確かに，例として挙げられているのは英語の特定の単語であり例文であるし，LOB コーパスは現在から見ればむしろ小規模なコーパスであって，データの量と多様性において限界があることは否めない。しかし，コーパスが言語学研究に有益であると言う論旨の大筋は一般言語学的なものであり，他の言語におけるコーパス言語学においても妥当することは明らかである。

リーチは滞日の期間中にこれと同様の趣旨の講演を何度か行ったようであり，したがって，日本の言語研究者は1980年代の前半にはすでにイギリスのコーパス言語学の成果に触れる機会があった。しかしながら，日本においてその直接的な影響が直ちに現れるということはなかったし，そもそもその当時においてリーチの講演の意義を真に理解できた日本の学者は多くなかったように思われる。実のところ，その講演を聴いたはずの筆者にも，それほど印象的なものとして記憶にとどまってはいない。また，この大会の概要を報告する近藤(1985)も，リーチの講演に触れた段落の中でその大筋を的確に紹介しながらも，その「資料体」がコンピュータ上のものであるという重要な事実には触れていなかった。

なお，この講演は，後に，全体の趣旨をほとんど変えない形で，Leech(1990)として論文にまとめられたが，日本のみならず海外においても，これまであまり引用されることがなかった。しかし，最近刊行された Teubert & Krishnamurthy(2007)に再録されたことに示されているように，この論文はコーパス言語学の発展を示す歴史的意義を有するものである。

このようにリーチの講演は当時の日本においてさほどの反響を呼ばなかったし，コーパス言語学一般への関心を高めるという結果をもたらしもしなかった。先端的な研究の紹介がこのような結果に終わってしまった理由をここで反省してみることは意味のあることであろう。その理由を完全に特定することはできないが，いくつかの複合的な理由を挙げることはできよう。

一つには，この講演の趣旨が，英語における特定の類義語の区別という，ある意味で瑣末な問題のみを扱ったものと誤解され，実際より価値を低く見積もられる可能性がなかったとはいえない。上述のように，Leech(1990)として印刷された際には題名が'The value of a corpus in English language research: A reappraisal'（「英語研究におけるコーパスの価値：再評価」）と変

更されている。リーチの元の講演が英語で行われ，Leech(1990)が日本で刊行されたものの英語で書かれていたこと，また，題名がより限定的なものになったことは，それを受容する層を一定程度に狭めてしまうことにつながったと考えられる。とりわけ国語学分野の研究者に対しての訴求力を減殺したであろうことは容易に想像できる。

しかしながら，この講演の影響が小さいものにとどまったことのより大きな理由は，聴く側にそれを受け入れる用意が整っていなかったところにある。1984年当時，16ビットパソコンが出始めてはいたものの，ハードウェア，ソフトウェアの両面で数メガバイトのデータファイルを扱うことは現実的ではなかったし，大型計算機の敷居は高すぎて，人文系の研究者には容易に近づけるものではなかった。それ以前に，コンピュータをワープロとして利用することはともかく，言語研究のデータを扱うということは多くの研究者にとってはコンピュータの利用法として想像の外であったと言えよう。

確かに，日本でも一部の言語研究者はこの時期にすでに言語研究へのコンピュータの応用を考え，実行していた。実際，日本のコンピュータを用いた計量言語学の歴史は古い。国立国語研究所は1966年にコンピュータを導入し日本語研究への応用を始め，文字や語彙（用語用字）の研究に一定の成果を挙げていた（国立国語研究所 1968-80，1970-1973など)。計量国語学会の発足はさらにそれに先立つ1957年のことであった。1980年代に入って，初期のパソコンでの試みとして，草薙(1983)の8ビットパソコン上のBASIC言語でかな表記日本語のKWICコンコーダンスの作成事例もある。英語に関しては，ブラウンコーパスに触れていた英語研究者はすでに存在していた（鈴木 1982）し，長瀬・西村(1986)につながっていくようなテキストの分析がイギリスの事例を参照して行われていた。とはいえ，これらの事例は，言語研究全体の中では例外的な存在であったと言えよう。コンピュータを使うことが自らプログラミングすることとほぼ同義であり，コンピュータ上での漢字の処理という日本語にとってより基本的なことが先決問題としてあったこの時期には，コンピュータで言語データを扱うというアイデア自体が，日本の言語研究者の大多数の間では未知の領域であった。

ただ，ここでさらに注意したいのは，リーチの研究方法には，単にコンピュータで大量の言語データを扱うという以上の意味があったことである。リーチの研究には次のような特徴があった。

・特定のテキストではなく，一言語の体系ないし運用のしくみ（ないしその特定部分）を記述の対象とする。

・孤立した語やその集合としての語彙目録の特徴づけでなく，テキストの中での個々の語の振舞いに関心をもつ。

・言語を複数の層にわけて捉え，それらの層の間での違いに関心をもつ。

これらの特徴は，リーチの意味論研究者としての関心の自然な延長として容易に理解することができるが，これ以降のイギリスにおけるコーパス言語学の諸研究においても主流をなすものであり，コーパスを利用する種々の研究方法のうちでコーパス言語学を特徴づける性質と見做すことができる。一方，これらの特徴は，当時の日本におけるコンピュータ利用言語研究にはあまりみ

られないようである。

このような研究方法が可能であったのは，リーチの依拠するデータであるLOBコーパスがそれを可能にする理論的基盤の上に載っていたからである。LOBコーパスは言語研究の目的をもってあらかじめ設計された，いわゆる「狭義のコーパス」であった（後藤 1995，2003）。具体的には，1961年の英語の書き言葉を母集団として設定し，15のジャンルごとにバランスをとって，印刷刊行された新聞，書籍，雑誌等から各約2000語のテキスト（の断片）をランダムに500集めて母集団を代表させるという設計であった。このため，リーチはLOBコーパスのデータをもとにして当該の語の語彙体系の中での位置づけを論じることができたのである。ただし，後述のように，LOBコーパスの設計は独創的なものではなく，ブラウンコーパスの設計に倣うものである（Johansson et al. 1978）。また，LOBコーパスが特定の年代のイギリス英語を真に代表するものであるかの議論はありうる。しかしそれは何をもってコーパスの代表性を保証するかという，より根本的なテーマにおいて論じられるべきものであり，本論の範囲では，代表性を正当化するための根拠を有していたことを指摘すれば十分である。

残念ながら，リーチの講演において，LOBコーパスのこのような性質は十分に説明されなかった。特定のテキストではなく，言語研究の目的に適合するようにとの意図をもって設計された均衡コーパスというアイデアは，当時の日本では一般には知られていなかったのであり，聴衆の多くはリーチが再評価を呼びかけたコーパスの性質について十分に理解できず，講演の理論的前提条件をそもそも捉えられなかった。このため，Leech(1984)はその時においてそれに本来ふさわしい影響力をもつことができなかったのである。

3. 英語コーパス言語学の発展要因

それでは英語を対象とするコンピュータ利用の言語研究は，なぜ早い時期に前節でみたような性質をもつコーパス言語学として成立しえたのであろうか。これもまた確実な答えを出せる問題ではないが，その要因を挙げることは可能である。

その第一に考えるべきことは，言語学の流れに沿っていたという点である。ここでコンピュータ以前のコーパスに基づく言語研究の流れを無視することはできない。周知の通り，二十世紀の半ばにアメリカで展開した記述言語学は，アメリカ大陸土着の言語を扱う必要から，研究者自身の直感や既存の辞書や文法書の助けに頼ることができない状況で，フィールド調査によって得た言語データにもっぱら依拠して特定言語の音韻体系や文法体系を記述する手法を開発した。このアプローチはコーパスに基づく研究に近い。この学派の中には，Fries(1952)のように，自分のネイティブスピーカーとしての直感を意図的に排除して，もっぱら実際に観察された発話データに依拠して英語を記述する試みもあった。この研究は現在でもコーパス言語学の直接の祖とみなされることがある。

一方，イギリスには，言語研究に限らず全般的に，経験主義を重視する伝統的な学問風土がある。その中で，1950年代からQuirkらによってSurvey of English Usageのコーパス作成プロジェクトが取り組まれていた。これはコンピュータ以前のコーパスとして代表的なものだが，80年

代まで続いて，約5000語を含むテキストのサンプル200（話し言葉と書き言葉それぞれ100）を集めてイギリス英語を代表させることになる（Svartvik & Quirk（eds.）1980）。このコーパスは，後にこの時代を代表する記述文法書である Quirk et al.（1985）のデータとして使われた。

最初のコンピュータ・コーパスとしてアメリカのブラウン大学で作られたブラウンコーパスにおいては，1961年のアメリカ英語の書き言葉を，15のジャンルにわけ，それぞれのジャンルにつけられた重みに応じて2000語のテキストの断片を総計で500集め，全体として100万語規模のコーパスを作るという，独自のアイデアをみせた（Francis 1965）。ブラウンコーパスがこのように言語の代表性を実現しようとしたことはこれ以後のコーパスに対して見本となった。また，著作権の処理を適切に行い，データを研究者の間で共有できるようにあらかじめ配慮していたことにも注意すべきである。このようにブラウンコーパスはコーパスの古典として位置づけられるにふさわしい性質を備えていた。ただ，二十世紀後半のアメリカの言語学はチョムスキー流の生成文法が主流となったため，そのままアメリカにおいて発展するには至らなかった。

イギリスの経験主義的な言語研究の流れは，ブラウンコーパスの直接的な影響を受け，そのイギリス英語版にあたる LOB コーパスを生み出し，この流れは，さらにハリデイらの言語理論とも関係をもちながら発展していく。したがって，コーパス言語学は当初から英語の文法構造ないし語彙構造の記述という目的があったと言える。ただし，コーパス言語学の成果が目に見えるようになるのは，1978年の LOB コーパスの完成からしばらくして，コンピュータの性能と利用の便宜がある程度まで整う1980年代になってからのことである。Leech（1984）は，したがって，現実の成果をもってコーパスの有用性を広く言語研究者にアピールする初期の呼びかけの一つという意味をもっていたのである。なお，corpus linguistics「コーパス言語学」という言い方が現れるのは1984年ごろのことであるが（齊藤他 2005：3），これはこのころに研究手法として確立されたことを意味しており，Leech（1984）の年代と符合するのは必ずしも偶然ではない。

その後，イギリスでは，外国人学習者向けの辞書の編纂を中心とする COBUILD Project（Sinclair（ed.）1987）の成功もあって，80年代末にはコーパス言語学は十分に地歩を固めることができ，Svartvik は1991年のコーパス言語学のシンポジウムに寄せた巻頭論文（Svartvik 1992）を “Corpus linguistics comes of age” と題するまでになった。1994年には，イギリス英語を代表する1億語規模の British National Corpus（BNC）が6機関の共同作業の結果，完成した。その後の英語コーパス言語学は，インターネットの普及などの状況の変化とともにその方向性を多様化させており，ウェブ上の言語データなど，従来の設計を重視するコーパスとは違う種類のコーパスを志向する動きも見られる。

英語コーパス言語学の発展に関しては，他にもいくつかの要因が考えられる。例えば，英語という言語が，その書き言葉の形態において，コンピュータで扱いやすい ASCII 文字でほぼ済むこと，単語をスペースないし句読点で区切られた文字列としておおむね定義できること，語形変化に乏しいことなど，言語内的に有利な要因もある。英語では，単純な文字列検索によって近似的には語などの言語的単位を検索でき，それによって一定程度にはおもしろい結果をえることができる。ここから，テキストデータの質を高め，量を増やし，また，検索プログラムの機能やイ

ンターフェースを洗練されたものにするなど改良の動機付けがはたらき，一層有益な結果を得ることにつながった。このような循環がはたらきやすかったのである。

さらに別の種類の要因として，1990年代には英語が「地球語」とも形容されるほどに他を凌駕する大言語となり，研究と教育における実践と応用が世界各地で広く行われたこと，またその結果として，学習辞典に端的に示されるようにビジネス上の利益にも結びつきやすいことなど，言語外的要因も決して小さいとはいえない。

これらの要因が複合的に関連しあって相乗効果を発揮しながら，英語コーパス言語学が早期に成立したものと考えることができる。

4. 日本におけるコーパス言語学

1984年のリーチの滞日時のような散発的な出会いはあったものの，コーパス言語学が日本で意識されるようになるのは，1990年代に入って英語コーパス言語学の成果が知られるようになってからである。1993年の英語コーパス研究会（のち英語コーパス学会）の発足を機にしてコーパス研究が急速に本格化する（齊藤他2005：7 - 8も参照）。この学会は歴史言語学や文学研究，英語教育への応用をも含めて，広くコーパス研究全体を包括する性質をもっているが，その中心はやはりイギリスのコーパス言語学の影響を強く受けた語彙や文法など共時的な言語研究にある。「コーパスを全面的に活用した初の英和辞典」であることを謳う井上永幸・赤野一郎編『ウィズダム英和辞典』（三省堂，2003）はこれらの活動の大きな成果の表れと言える。

1990年前後にはパーソナルコンピュータの普及が本格化し，それを日本語研究へ応用する試みが各所で始められる。しかし，多くの場合，計量国語学との連続性を欠く形で試行錯誤的に行われた（後藤 1995）。また，英語コーパス研究会の成立は日本語学の分野にとっても一つの刺激であったと思われる（伊藤 1994）が，理論的な相互の影響関係はあまり認められない。

最も惜しいのは，日本語研究の分野ではコーパスの設計に関する議論が十分に行われず，このこともあって人文系の言語研究者の間で共有できる狭義の日本語コーパスが長い間存在しないままにされたことである。筆者は後藤(1995)において，日本語コーパスの方向を二通り考えた。一つは綿密な設計に基づくコーパスであり，もう一つは全体のバランスはある程度は度外視してもできるだけ大きな量のテキストの集積であるが，後者の場合であっても，できるだけ多様なタイプのテキストが含まれるようにすべきであることを指摘した。

残念ながら，前者の方向での進展はほとんど生まれず，後者の方向でも，量的な拡大はあったものの，十分な多様性を確保するための方法論は確立されていない状況にある。そのため，日本語のコーパス研究では，新聞記事や文学作品など，主として入手の便宜を理由として選択された広義のコーパスを利用する研究手法が続いてきている。全体として散発的であることは否めず，英語の場合に比べられるような大きさのインパクトを言語研究やその応用分野に与えてはこなかった。この意味で，いまだ日本語コーパス言語学が成立しているとは言いがたい。

このようななかで，国立国語研究所が「現代日本語書き言葉均衡コーパス」の構築に着手したことの意義は大きい。とりわけ意味があるのは，このコーパスが現代日本語を代表するようにと

の明確な設計に基づいていること，および，それが公開される予定であること，である。言語には無限の生産性があるので，いかに大規模なコーパスといえども完全に代表することはできないが，その近似値を知るための大きな手掛かりを提供してくれる。他のテキストデータを使う場合にも，そのテキストの性質を客観的に把握するための比較の基準として使うことができる。また，データの公開によって，多くの研究者がコーパスの便宜を実際に享受できるようになるだけでなく，研究結果を他の人が追試することを可能にもし，成果をより客観的なものにすることになる。

このようにして近い将来に日本語のコーパスが広く使われるようになることは極めて望ましいことである。それを十分に活用するためには，それが存在するだけでは不十分であり，利用者の側にその活用に必要な知識と技能を得ようとする主体的な努力が要求される。コーパスは手軽に情報を得ることのできるブラックボックスではないのであり，その性質を十分に理解した上で扱わなければ意味のある結論には結びつかないからである。コーパスを活用するには，言語とコンピュータの両方の分野に関しての知識が必要である。

言語研究の手段としてコーパスを利用するのである以上，言語学の考え方が予備知識として必要なのは当然のことである。ただし，コーパスを適切に扱って，コーパスから得られたデータから意味のある情報を読み取るためには，やはりそれなりの手法を身につける必要がある。英語に関しては，Sinclair(2003)のような，コンコーダンスの実例を豊富に挙げて，それを例題として，手順を経ながら語義分析を進める過程を訓練するためのテキストがある。この本の解答例は必ずしもすべて納得のいくものではないが，COBUILD のプロジェクトを先頭に立って推進した辞書編集者の描くコンコーダンス分析の手順はなんといっても大いに参考になる。また，スタッブズ(2006)は訓練用ではないが，分析の記述が具体的であり，それに近い性質をもっている。残念ながら，日本語には，このようなテキストはまだないため，コーパス言語学を実践する研究者は，自力でその方法を習得して行かざるをえない。

これに加えて，コンピュータを言語の研究に利用するのであれば，コンピュータが文字およびテキストをどのように扱っているかについての基礎知識が欠かせない。しかし，これに関する知識は言語研究者の間にはそれほど普及していないのが実情である。現在では，コンピュータの利用自体は普及しており，ワープロや電子メールを扱うのに苦労することはほとんどなくなっている。しかし，言語研究者の間でも，その利用法は，往々にして，一般人と同程度に表面的なものにとどまっている。

例えば，文字集合やコードの知識はテキストを扱う際に文字通り基礎となる知識である。とりわけ日本語のような，複雑な書記体系をもち，文字使用において習慣的に高い自由度を許容している言語を扱う場合には，その知識が不可欠であることは当然である。今後，日本語においても Unicode の使用が広がっていくことは間違いないが，国内の規格として従来広く使われていた JIS コードで書かれたテキストも当分は流通し続ける以上，この二つのコード体系およびその関係を知らずには済ませられない。しかしながら，言語の研究者にとって，その知識を系統的に習得する機会は多くない。それを得ようとする意識的な努力が求められるのである。

また，コーパス研究に役立つ使いやすいソフトウェアは今後ますます増えていくであろう。しかし，自分で入手したテキストを自分なりにコーパス（ないし，その一部）として使いたいという希望を持つこともまた自然なことである。そのような場合，テキストを編集する作業を行うことになるが，正規表現による検索や置換を行えるツールを使えば，一々手作業で行うのに比べて，その編集作業は劇的に軽減される。既存のデータやソフトウェアに全面的に頼るのでなく，自分でさまざまな工夫をしようと思うのであれば，やはりさまざまなテキスト・ツール類の使用法を自分で習得しなければならないのである。

5. 語彙論への応用の試み

コーパス言語学は，語彙，文法，言語変異など，実際のデータからの根拠を必要とする言語研究の多くの分野において有効である。しかし，理論的には可能であっても，コーパス言語学の方法の蓄積に乏しい日本語などの言語においては，コーパスができたからといって，コーパスの利点を理論的に可能なすべての面にわたって発揮することは，直ちには難しいと考えるべきであろう。

例えば，文法タグの付けられたコーパスはプレーンなコーパスに比して情報量が多いため，有用性が高くなることは疑いない。ところで，文法タグはあらかじめある程度の細かさで用意されているとはいえ，特定の研究者が関心をもっている個別の文法現象と形式的に直接対応しているとは限らない。研究者が自分が必要とする情報をコーパスから得るためには，単にソフトウェアを操作して特定のタグのついた部分を検索するだけでは足りず，なにがしかの試行錯誤的な工夫が必要になる事態が発生しがちである。しかし，日本語研究者の大部分は，これまで文法タグの付いたコーパスを扱った経験がなく，そのような工夫には不慣れである。コーパスから自動的に結果が得られるという過大な期待を抱くならば，それは往々にして裏切られ，コーパスへの不信にもつながることになりかねない。

日本語の研究にとって，より早期に効果が期待できるのは，形式的な同定が容易な語彙のレベルであろう。例として「喫緊」という語を取り上げてみる。これは，辞書によって多少の違いはあるが，おおむね「差し迫っていて，非常に重要なこと」（『明鏡国語辞典』）のような語義によって説明される。この語義は妥当であるように思えるが，この語は現代語においては，用法が極めて制約されており，国語辞典に記されない特徴をもっている。

筆者の所有するデータからは，生起するテキストのジャンルの制約（位相上の制約と考えてよい）とがあること，および「課題」以外の語が後続することがほとんどないというコロケーション上の制約があることが示される。筆者のデータにはこの語は合計で153回現れるが，そのうちの96例は各省庁が刊行した白書に現れるものである。いわゆる「お役所ことば」であるが，この語が現れるのは，白書やそれに類する公的文書のほかに，シンクタンクの報告書，業界団体の広報および新聞記事のうち政治欄や社説などに及んでおり，官庁の文書よりはいくぶん広い。なお，このことは，一口に新聞記事といっても性質の異なるものが含まれていることを示唆する。「喫緊」はその他のジャンルのテキスト，とりわけ小説などにはほとんど現れない。

この語は，その生起のほとんどを占める137例で「喫緊の課題」という結びつきで現れるという際立った特徴をもっている。ほかに「喫緊の政策課題」が3例，「喫緊の国家課題」および「喫緊の，自己のアイデンティティにかかわる課題」が各1例あり，「喫緊」と「課題」との緊密な関係は明らかである。ほかに「テーマ」や「対策」とともに使われた例もあるが，それぞれ少数であり，目立たない。このような強い結びつきを示す性質は，「喫緊」の類義語とみなせそうな「緊急の・重要な」などと「課題」の類義語である「問題・急務」などの間にはみられないものであり，この語に特徴的なものである。一部の辞書は次のような用例を挙げている。実際の用例を記録することを目的とする大辞典における用例としては意味のあることだが，この用例はこの語の使い方をよく例証するものというより，当該の文章の古さないし文体的な特殊性を示すものと考えるべきであろう。

(1) 鶴川の死は父の死にもまして，私に喫緊の問題とつながりがあると思われたからだ　三島由紀夫・金閣寺　『学研国語大辞典』

(2) 真を極むるの道に於て―必須　真善美日本人　雪嶺　『大辞林』

上でみたような「喫緊」の性質はある程度まで内省によって知ることができるが，実際の生起における偏りを明確に知るためにはコーパスを参照することが必要である。

しかしながら，ここで使ったデータは，狭義のコーパスではなく，筆者がたまたま収集することのできたテキストの集合である。筆者が市販のテキストを個人的に収集したものであり，事前に全体を設計したものではない。これにはいくつかの決定的な欠点がある。これらはそもそも無原則的に集められたものであり，さまざまな位相の間での違いを印象以上に述べることが難しい。ここで言えることがどの程度まで現代日本語に対して一般化できるかは明らかでない。用例の実数を挙げてはみたものの，その数字にどれほどの意味があるのか，疑わしい。「ほとんど……ない」，「多い」，「目立つ」などの印象的な，曖昧な表現とあまり変わらない。さらに，データは研究に利用するための特別の著作権処理をしていないため，個人的な利用は可能であるが，複製することは許されていない。そのため，例文を直接引用することも避けた。したがって，他者が同じデータを使って検証することができない。このようなことは，研究の基礎のデータとして用いるには，本来望ましくない形態である。

コーパスが整備されることによって，ここで行ったような記述がより精緻化され，積み重ねられていけば，語彙項目間に見られる関連や文法現象との関連に対するより深い理解につながることが期待でき，さらには語義のより深い分析や，文法や語用論の面のコーパス言語学も次第に整うであろう。現状においては，コーパスを用いた日本語の語義分析はその精緻さにおいて英語コーパス言語学のスタッブズ(2006)の域には到底達していないし，Leech(1984)にも及ばないと言わざるをえないが，日本語の大規模な均衡コーパスの整備は現在のような事態を大幅に改善するきっかけとなるであろう。

6. まとめ

英語コーパス言語学は，21世紀にはいってその方向性を多様化させており，ウェブ上の言語デ

ータなど，従来の設計を重視するコーパスとは違う種類のコーパスを志向する動きも見られる。しかし，それらは均衡コーパスの成果の上に立ち，言語の一層の多様性を見たいという動機によるものであり，均衡コーパスの存在意義を否定するような性質のものではない。

国立国語研究所が2006年度から5年の計画で「現代日本語書き言葉均衡コーパス」の構築に着手したことは日本語コーパス言語学の発展にとって大きな意義をもつことと言える。ただし，それを十分に生かすためには，利用者の側にもそのための技術を習得する主体的な努力が求められる。これは必ずしも楽観できることではないが，将来の可能性に期待したい。

参考文献

伊藤雅光(1994)「数理的研究」『国語学』177，121-138，国語学会

草薙裕(1983)『コンピュータ言語学入門』大修館書店

国立国語研究所（1968-1980)『電子計算機による国語研究Ⅰ-Ⅹ』(国立国語研究所報告）秀英出版

――― (1970-1973)『電子計算機による新聞の語彙調査Ⅰ-Ⅳ』(国立国語研究所報告37，38，42，48）秀英出版

後藤斉(1995)「言語研究のデータとしてのコーパスの概念について　―日本語のコーパス言語学のために―」『東北大学言語学論集』4，71-87，東北大学言語学研究会

―――(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』4月臨時増刊号「コーパス言語学」，6-15，明治書院

近藤達夫(1985)「日本言語学会第89回大会報告」『月刊言語』14(1)，254-255，大修館書店

齊藤俊雄他(2005)『改訂新版　英語コーパス言語学』研究社

鈴木英一(1982)「ブラウンコーパスへの招待」『月刊言語』11(10)，113-119，大修館書店

スタッブズ，マイケル(2006)南出康世・石川慎一郎監訳『コーパス語彙意味論』研究社

長瀬眞理・西村弘之(1986)『コンピュータによる文章解析入門―OCPへの招待―』オーム社

Francis, W. Nelson(1965) A standard corpus of edited present-day American English, In G. Sampson & D. McCarthy(eds.)(2005) *Corpus linguistics: Readings in a widening discipline*, 27-34, London/New York: Continuum International.

Fries, Charles C.(1952) The structure of English, In G. Sampson & D. McCarthy(eds.)(2005) *Corpus linguistics: Readings in a widening discipline*, 9-26, London/New York: Continuum International.

Johansson, Stig, Geoffrey N. Leech, & Helen Goodluck(1978) *Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers*, Oslo: Department of English, University of Oslo.

Leech, Geoffrey N.(1984) The value of a corpus in linguistic research: A reappraisal, 第89回日本言語学会大会公開講演（東北大学大学院文学研究科言語学講座所蔵録音テープによる).

――― (1990) The value of a corpus in English language research: A reappraisal, 筧壽雄教授還暦記念論集編集委員会編『ことばの饗宴―筧壽雄教授還暦記念論集』くろしお出版.

Quirk, Randolf, Sidney Greenbaum, Geoffrey N. Leech, & Jan Svartvik(1985) *A comprehensive grammar of the English language*, London: Longman.

Sinclair, John M., ed.(1987) *Looking up: An account of the COBUILD project in lexical computing,*

London: Collins.
——— (2003) *Reading concordances*, London: Pearson Education.
Svartvik, Jan (1992) Corpus linguistics comes of age. In Jan Svartvik (ed.) *Directions in corpus linguistics: Proceedings of Nobel Symposium 82*, Berlin/New York: Mouton de Gruyter.
Svartvik, Jan & Randolph Quirk, eds. (1980) *A corpus of English conversation*, Lund: C W K Gleerup.
Teubert, Wolfgang & Ramesh Krishnamurthy, eds. (2007) *Corpus linguistics (Critical concepts in linguistics)*, 1, London: Routledge.

（投稿受理日：2007年8月3日）

後藤　斉（ごとう　ひとし）
東北大学大学院文学研究科言語学研究室
980-8576　仙台市青葉区川内27番1号
gothit@sal.tohoku.ac.jp

Corpus linguistics and Japanese language studies

GOTOO Hitosi
Tohoku University

Keywords
corpus, corpus linguistics, Balanced Corpus of Contemporary Written Japanese

Abstract

Linguistics in Japan has failed to develop corpus-based language studies into corpus linguistics, inspite of the long history of computer-based mathematical linguistics dated from the 1960s and sporadic contacts with English corpus linguistics since the 1980s. This is contrastive to the situation in Britain, where corpus linguistics has been established since the early 1980s, with grammatical and lexicological studies as main foci of interest.

It is noteworthy that there is no Japanese corpus, available to researchers, which could be safely claimed as representative, so that researchers are now obliged to use a haphazardous collection of electronic texts as a corpus. Usefulness of such a corpus is evident, as is shown in a tentative case study, but inevitably limited. A representative corpus would serve better to linguistic research.

The project of Balanced Corpus of Contemporary Written Japanese, now being undertaken by the National Institute for Japanese Language, is expected to fill the need and this is evidently welcome. It should be noted, however, that, in order to gain full advantage of a corpus, users will have to make efforts to acquire knowledge on techniques and basic facts in text processing.