探索的コーパス言語学のための覚書

A Note for Exploratory Corpus Linguistics

石井　正彦
ISHII Masahiko

キーワード：確認的／探索的，理論モデル主導型／データ主導型，コーパス検証型／コーパス駆動型，語彙調査，連結パターン

要　旨

統計学とその応用分野で生じた，データによって仮説を検証する確認的な（理論モデル主導型の）アプローチから，データによってデータそのものを説明する探索的な（データ主導型の）アプローチへの展開が，同じくコーパス言語学でも生じていることを跡づける。その上で，今後，後者のアプローチを採る「探索的コーパス言語学」と呼ぶべき日本語研究が大きく展開していく可能性を示し，そのための方法論として，(1)連結パターンの探索，(2)全文コーパスの利用，(3)多様なコーパスの開発，(4)探索的データ解析の適用，の４点を提案する。

１．はじめに

表題にいう「探索的コーパス言語学（Exploratory Corpus Linguistics)」とは，コーパス言語学において，これまで主流であるところの「確認的なアプローチ」の研究に対し，それとは異なる「探索的なアプローチ」による言語研究を標榜して，筆者が仮に設けた名称である。ここで,「確認的（confirmatory)」対「探索的（exploratory)」とは，統計学の探索的データ解析における用語を借用したもので,「理論モデル主導型（theoretical model oriented)」対「データ主導型（data oriented)」という用語で呼ばれることもある。その詳細は後述するが，ごく簡単に言えば，確認的な（理論モデル主導型の）アプローチとは「データによって仮説を検証する」ものであり，探索的な（データ主導型

の）アプローチとは「データによってデータそのものを説明する（データ自身に語らせる）」ものである。本稿では，統計学とその応用分野において生じた理論モデル主導型からデータ主導型への展開がコーパス言語学においても生じていることを跡づけ，その上で，今後，「探索的コーパス言語学」と呼ぶべき日本語研究，すなわち，コーパスを用いたデータ主導型の日本語研究が大きく展開していく可能性があることを示して，そのための方法論的な検討を行う。

２．大規模コーパス言語学

2011 年に国立国語研究所が構築・公開した『現代日本語書き言葉均衡コーパス（BCCWJ）』は，日本で最初の本格的な大規模均衡コーパスとして，日本語コーパス言語学の歴史を画するものといえる。田野村忠温は，これについて以下のような賛辞を呈しているが，筆者もまったく同感である。

> BCCWJ をただ所与の資料として利用するという意識でいては分かりにくいが，その開発は背後の膨大にして綿密な計画と作業に支えられており，BCCWJ の完成は偉業の名にも値する。その開発に当たった関係各位に対しては敬意と感謝の念を禁じ得ない。BCCWJ を質・量の両面において凌ぐ日本語のコーパスが近い将来に再び作成されることはないであろう。（田野村 2014：119）

こうした大規模コーパスの実現は，コーパス言語学に限らず，実証的な言語研究全体にとって直接の大きな利点となる。BCCWJ の有力な後援者であり理解者であった宮島達夫は，次のように述べている。

> 単にコーパスを利用して言語現象をしらべた，という研究を「コーパス言語学」とよぶ必要はない。コーパスの第１の価値は，膨大な用例の量にある。これからの研究には，当然それを利用すべきだが，それは「コーパス言語学」でも「用例言語学」でもない。大量の例文をしらべることは，まさに言語学の王道，限定語なしのザ・言語学だ，というのが，わたしの立場である。（宮島 2007：41）

一方で，コーパス言語学の立場からは，大規模コーパスの実現は言語研究に「分析者が科学的手法に則って言語研究を行えるようになる」という重要な利点をもたらす，ともいわれる（マケナリー＆ハーディー2014：21）。ここでいう

「科学的手法」とは，自然科学の分野を典型としていわれる「仮説形成検証過程による真理探究の方法」（冨田・三輪 2002：501）をさすものと思われる。田野村(2000：193)が「大規模なコーパスから得られる用例は用法の現実を忠実に反映して網羅的であり，内省による予想をはるかに超えて多様である」と言うように，大規模コーパスは網羅的かつ多様であり，その網羅性・多様性が，仮説形成（発見）の契機をより高い確率でもたらし，また，仮説検証のための反証可能性や再現可能性といった特性・資格を大規模コーパスに与えていると考えられる。これにより，分析者は，自らによる仮説の形成と（他者による仮説も含めた）その検証とを連続的・循環的に行って真理に近づくという，科学的手法を獲得・駆使することができるのである。

３．統計学における方法論的展開

ただし，コーパスを用いた言語研究が，すべて，大規模コーパスによる仮説形成検証過程の科学観によって一元化されるわけではない。そのことを述べるにあたって，筆者は，まず，20 世紀半ば以降の統計学の方法論が「理論モデル主導型（theoretical model oriented)」といわれる方法から「データ主導型（data oriented)」といわれる方法へと大きく展開したと説かれていることに注目したい。その概要を吉田編(1995)によってまとめると,以下のようになる。

> 現代の統計学は，20 世紀初頭，理論モデル主導型の数理統計学として成立した。ここで，理論モデル主導型とは，収集されたデータが確率的な要因に支配される場合に，データの背後にデータの発生のしかたを規定する確率的なモデルを想定し，それをもとに確率モデルの推定や検定を行うという「統計的推測」にもとづくことをいい，数理統計学はこの統計的推測の数理的な枠組みを研究する学問分野として完成された。その後，1950～60 年代には数学的形式化がさらに進み，数理統計学は数学理論としての自己完結化への道を歩むが，一方では，その形式性，抽象性のゆえに科学的分析を保障する方法として社会科学，自然科学を問わず応用分野の多様化が進んでいく。しかし，これらの応用分野で得られるデータは，実験データであったり観測データであったりと，分野によってさまざまな特徴・性格をもっていたため，確率分布（確率モデル）を想定して標本から母集団

パラメータを推測する従来の理論モデル主導型の方法では十分対応できないという反省が数理統計学の中から生まれ，数学的前提を仮定することなくデータの構造や特徴を的確に要約・再現する（＝データに語らしめる）ための「データ主導型」の方法が提唱されてくる。その端緒となったのが，アメリカの統計学者テューキーによる「探索的データ解析」の考え方である。データ主導型の方法には，このほかにも，多変量解析や数量化理論などさまざまなものがあるが，1980 年代以降になると，コンピュータの急速な進化も相まって，これらデータ主導型の数理統計学が大きなうねりとなって進展している。（吉田編(1995)「第 2 章　統計学の歴史」「第 15 章　数理統計学の現状」から要約）

ここで，理論モデル主導型とデータ主導型との違いをより具体的に理解するには，後者の端緒となった探索的データ解析の考え方が参考になる。探索的データ解析（Exploratory Data Analysis；EDA）とは，推測統計学に代表されるような，統計的仮説を検証するための「確認的データ解析（Confirmatory Data Analysis；CDA)」に対して，限られたデータから何らかの仮説的な情報を探り出すことを目的として，アメリカの統計学者テューキー（J.W.Tukey）が中心となって主に 1970 年代に開発した一連のデータ解析の手法や考え方をさす。そこでは，CDA のように標本から母集団を推定するということ，すなわち，「部分から全体への推論を明確な手段をもって行う」理論モデル主導型の方法は採らず，どのようなデータでも対象として，「データに含まれる混沌とした情報を整理し，データについての洞察を得，データの構造を探り，そうすることによってデータの説明を行おうとする」「いわばデータそれ自身に語らせる」（渡部ほか 1985：170）ことが行われる。テューキー自身の比喩によれば，理論モデル主導型の CDA が得られた証拠から犯罪を判定し量刑を確定する検事や裁判官のような仕事であるのに対して，データ主導型の EDA は証拠を収集して立件していく刑事や警察官のような仕事であるという（吉田編 1995：104)。

統計学における理論モデル主導型からデータ主導型への展開は，先の要約にもあるように，応用分野の拡大に伴い，対象とするデータが実験データやそれを擬制したものから観測データへと拡大し，質的にも量的にも多様なデータを処理しなければならないという状況に直面したことが主な要因である（このほ

か，とくに 1980 年代以降，人々の想像を超えるコンピュータの発展によって，大量かつ複雑なデータを適切に集約して「データに語らしめる」手法が開発されたことも，その背景にある)。このことは，こうしたデータ主導型への展開が，統計学を科学的探究の方法として採用する多くの応用分野で並行的に起こっていることを意味する。筆者が統計学における理論モデル主導型からデータ主導型への展開に注目するのは，まさにコーパス言語学においても基本的には同様の展開が生じていると考えるからである。ただし，その展開の過程はいささか複雑であり，多少の説明を要する。

４．コーパス言語学における方法論的展開

この点について，英語を中心とする欧米のコーパス言語学で注目されるのは，「コーパス検証型（corpus-based[1])」対「コーパス駆動型（corpus-driven)」という考え方である。研究史的には，従来からのコーパス検証型に対して，新たにコーパス駆動型のアプローチが定置されたわけだが，この区別を最初に提唱したとされる Tognini-Bonelli(2001)は，これらの用語（概念）について，以下のように説明している（日本語訳は，マケナリー＆ハーディー2014：222-223 による)。

> コーパス検証型言語学という用語が示すのは，大規模コーパスを使用して言語研究の知見を得ることができなかった時代に提唱された言語理論や言語記述について，もっぱらそれらを説明・検証・例証する目的でコーパスを用いる方法論である…コーパス検証型研究手法を取る言語学者は，理論とデータの関係について「自信」を持っている。彼らは，もともと基本的に正しいと考えている言語モデルや言語記述を持ち込み，それらが提唱する範疇に基づいてコーパス分析を行ない，それに沿ってデータをふるいにかけるからだ。
>
> ――――
>
> コーパス駆動型研究手法では，言語学者の責任はデータ全体を一体的に扱うことにあり，コーパスから得られる言語的証拠に関して，包括的言語記述を行うことが目指される。それゆえコーパスは，既存の言語理論を支持するための単なる用例集以上のものとして，また，確立された既存の理

論体系を拡張しうる手段以上のものとしてとらえられるのである[2)]。この手法では，コーパスに見られる言語的証拠と完全に整合し，かつ，それを直接的に反映した理論的説明がなされる…理論が証拠から独立して存在することはないし，理論化に至る一般的な方法論上の手続きも明白で，観察から仮説構築へと進み,一般化を経て理論的説明との統合に至るのである。

これによれば，まず，コーパス検証型は，以下に述べるように，理論モデル主導型にほぼ対応しているように思われる。

現在のコーパス言語学は,BCCWJのような大規模均衡コーパスを一つの典型・理想形としている。先述したように，こうしたコーパスは基本的に標本－母集団モデルすなわち理論モデル主導型の考え方にもとづいて構築されている。この，標本（コーパス）から母集団を推定する（統計的推測）という行為の内実は,データの背後にデータの発生のしかたを規定する理論モデル(確率モデル)を仮定し，その確からしさを検証すること，すなわち，データをもとに統計的仮説を検証すること（統計的仮説検定）である。要するに，理論モデル主導型の，標本から母集団を統計的に推測するという特徴は，データによって仮説を検証する（確認する）こと，とも言い換えられるわけである。

理論モデル主導型の特徴をこのようにとらえれば，それは上にみたコーパス検証型の説明，すなわち,「コーパスデータを用いて，何らかの理論や仮説，とくに現在までにすでに確立されている理論や仮説を考察・証明・否定・精緻化するもの」（マケナリー＆ハーディー2014：8）という説明にもあてはまるものである。もちろん，ここでいう「仮説検証」の意味合いは「統計的仮説検定」に限られるものではないが，重要なのは，コーパス検証型もまた，理論モデル主導型と同様に,「データによって仮説を検証する」ということを特徴とするアプローチだ，という点である。

次に，コーパス駆動型について考える。この“corpus-driven”という用語の意味するところについては,コーパス言語学者の間にさまざまな見解(の相違)があるようである。マケナリー＆ハーディー(2014：223)は，上の Tognini-Bonelli(2001)の説明を引用した後，次のように述べる。

Tognini-Bonelli はこのように述べているが，コーパス検証型とコーパス駆動型の区別は，単純な二項対立がしばしばそうであるように，実際に

> はもう少し流動的なものである。コーパス検証型の研究者は既存理論の枠組みを使用しようとするが，その枠組みを実際のデータに適用してみた結果，うまくいかないとわかると，いわゆるコーパス駆動型と呼びうるような手法で元の枠組みの修正を行う。こうした過程は循環的なもので，言語学者全般,とくに計算言語学者は,以前よりこのことをよく理解していた。

この記述からは，コーパス駆動型というものを，コーパス検証型との単純な二項対立物と位置づけ，典型的な科学的手法としての仮説形成検証過程において，仮説検証とともに連続的・循環的に行われる仮説形成の方法を意味するものとする,マケナリー&ハーディー(2014)の解釈が読み取れる。また,同書は,“corpus-driven”という用語が「コーパスの生データを用いた帰納的・ボトムアップ的研究のすべてを指すものとしても使われている」とか,「コーパスデータを用いた方法論を採用しているという理由だけで」使われることもあるとかと述べ，さらに「探索的に，つまりは厳密に設定された仮説を持たずにコーパスデータを扱う（こと)」とするGilquin & Gries(2009：10)の見解をも紹介して，コーパス駆動型という用語が「実際に使用する上で明瞭性を欠いている」としている（マケナリー&ハーディー2014：223-224)。

しかし，私見では，コーパス駆動型を「元の枠組み，すなわち仮説がうまくいかないとわかると，その修正を行う」といった，仮説検証過程と連続的・循環的に行われる仮説形成過程と断じることは適当でないように思われる。先のTognini-Bonelli(2001)の「この手法では，コーパスに見られる言語的証拠と完全に整合し，かつ，それを直接的に反映した理論的説明がなされる…理論が証拠から独立して存在することはない」という説明は，コーパス駆動型が「データによって仮説を検証する」ものではなく,「データによってデータそのものを説明する（データ自身に語らせる)」ものであることを示しているようにも思える。いずれにせよ,「データに語らせる」というアプローチは，それ自身は検証されることを前提としていない。もちろん,そのようにして得た知見や洞察が,結果的に仮説として扱われ，検証過程に付されることはあっても，最初から仮説形成検証過程に組み込まれているとか，それを前提として存在しているとかと考えることは適切ではない[3)]。その意味では,マケナリー&ハーディー(2014)が紹介するGilquin & Gries(2009：10)が，先の引用の直後に,

> In other words, they seem to favor a corpus-driven approach, which ‘lets the data speak for themselves’. （言い換えれば，彼ら（＝コーパス言語学者）は，「データに自分自身を語らせる」コーパス駆動型アプローチを好んでいるように思える。）

と述べていることは示唆的である。このとらえ方は，コーパス駆動型が「データに語らせる」ことを特徴とするデータ主導型とその基本的な考え方において共通していることを示すものではないだろうか。それは，マケナリー＆ハーディー(2014)の意図に反して，コーパス駆動型に対する Tognini-Bonelli(2001)と Gilquin & Gries(2009)の理解が，少なくともこの点においては，同様のものであることを示しているようにも思われる。

以上のように，英語を中心とする欧米のコーパス言語学では，「コーパス駆動型」の内実に曖昧な点が残るものの，おおむね，理論モデル主導型に対応するコーパス検証型からデータ主導型に対応する（可能性の高い）コーパス駆動型への展開という，統計学やその応用分野と同様の動向が看取できるように思われる。

５．日本語コーパス言語学の現状

日本における（日本語の）コーパス言語学は，戦後に創設された国立国語研究所の一連の語彙調査を源流とするが，そこでは，基本的に，標本－母集団モデルに基づく推測統計学の方法，すなわち，理論モデル主導型の方法が採用された。先述のとおり，推測統計学は，標本をもとにその背後にある母集団の特徴（確率モデル）を推定・検定するのだが，標本データについて多量であることを前提にしないため，小規模な標本コーパスを手作業で作成すればよく，コンピュータのなかった時代でも実行できたからである。その代表といえる「雑誌九十種の語彙調査」は，1956 年 1 年間に発行された 90 種類の雑誌本文に用いられた延べ約 1.4 億語（β単位，推定）を母集団とし，そこから（無作為に）抽出した延べ約 53 万語を標本とするもので，手作業で行った語彙調査としては現在に至るまで世界最高水準のものとされている。

ただ，標本から母集団を推定するといっても，統計的に意味のある精度で推定できるのは標本に繰り返し現れる一部の語に限られてしまう。雑誌九十種調

査でも，母集団使用率の推定値（95%信頼区間）とその推定精度が付された語彙表（こうした語彙表自体，世界的にも例がないと言われている）が作成されたのは，標本使用度数が（全体で）50 以上の約 1200 語だけであったし，各語彙表に納められたのは標本使用度数 7 以上の 7200 語だけだった。要するに，53 万語の標本データをもとにその二百数十倍にあたる 1.4 億語の母集団の様子を推定するといっても，ある程度の確からしさをもって推定できるのは母集団の（もちろん中心部ではあるが）ごく一部分ということになってしまうのである。

その後，国語研究所は「新聞三紙の語彙調査」（1966 年）でコンピュータを導入し，延べ語数約 300 万語（短単位）の標本コーパスの作成をめざすが，当時のコンピュータの性能の限界（と日本語処理の難しさ）から，語彙調査に必須の単位分割（形態素解析）や同語異語判別を自動化することができず，語彙調査としては「失敗」に終わる[4)]。

それから半世紀近く，同じ国語研究所といっても，もはや語彙調査という枠組みではなく，コーパス言語学という新しい枠組みにおいて，BCCWJ は構築・公開された。そのデータ規模は延べ 1 億語（短単位）を上回るとされ，雑誌九十種調査のざっと 200 倍，同調査で母集団とされた 1.4 億語に迫る規模であり，統計的推測の精度は格段に向上しているはずである。同時に，均衡コーパスとして，雑誌だけでなく，書籍・新聞・白書・ブログなど異なるレジスターからもテキストを集め，書きことばの多様な変種とその実態をとらえることが目指されている。同コーパスの『利用の手引　第 1.1 版』[5)] によれば，BCCWJ 設計の基本方針の第一には次のようにある。

（1）現代日本語の縮図となるコーパス

従来，国立国語研究所が行ってきた語彙調査の手法を生かし，コーパスがその母集団の統計的な縮図になり，母集団に対し代表性（representativeness）を持つように設計する。これにより，母集団における言語的諸特性の分布が過不足なく表現できることになり，データの信頼性を高めることが出来る。

BCCWJ は「母集団の統計的な縮図」としての「標本」コーパスであり，その大規模性・均衡性にもとづく「代表性」によって統計的推測の精度すなわち「信頼性」を高めているのである。このことは（日本語のコーパスとして）画期的

であるが，しかしその基本的なアプローチが標本から母集団を推測するという理論モデル主導型のそれであることは明らかであり，それはまた半世紀前の語彙調査と基本的に変わらないものである。

では，日本語のコーパス言語学におけるデータ主導型アプローチへの展開とは，どのようなものだろうか。ここでも，その源流といえるものが国語研究所の語彙調査の中にある。いま一度，語彙調査の歴史に戻れば，「新聞三紙の語彙調査」の後，1970 年代後半から 80 年代にかけて，国語研究所は「高校・中学校教科書の語彙調査」にとりくむのだが，ここでは標本調査をやめて，高校・中学校の理科・社会科計 9 科目の教科書をそれぞれ 1 冊ずつ調査対象とする全数調査，すなわち，各教科書の本文を（標本抽出せず）すべて収める「全文コーパス」を作成してそれによる調査を行っている。そのねらいは，現行のすべての教科書の語彙を母集団としてその全体像を知ることよりも，特定の教科書の本文からその専門知識体系を記述するための語彙をすべて取り出し，その構造を明らかにするとともに，それによる文章構成の分析までを行うことにあったという（土屋 1989：2)。これは，データによって仮説を検証しようとする理論モデル主導型のアプローチでは少なくともなく，たとえ限られたデータであってもそのデータから構造やパターンを探り出そうする点で探索的であり，データによってデータそのものを説明しようとするデータ主導型のアプローチといってよいものである。

その後，国語研究所では，こうしたデータ主導型の全数語彙調査が行われることはなかった（わずかに「テレビ放送の語彙調査」（1989 年）が，その分析段階で探索的データ解析の手法を導入しているにとどまる)。しかし，1990 年前後から，新聞を中心に各方面の印刷用データがほぼそのまま「簡易コーパス」（マケナリー＆ハーディー2014：16）として利用できることも多くなり，新聞について言えば，「新聞三紙の語彙調査」で母集団とされた朝日・毎日・読売 3 紙の 1 年分に相当するテキストデータも現在では簡単に購入・利用できるようになった。これによって，「限られたデータ」どころか，かつてなら母集団と想定されるような規模のデータがまるごと調査できるようになり，少なくともこうしたコーパスを利用する限りは，標本から（統計的に）推測するという理論

モデル主導型の方法を採る必要がそもそもなくなりつつある。このような大規模な「全文コーパス」の実現は，日本語のコーパス言語学においても，理論モデル主導型からデータ主導型への展開をより強く推進する力になるものと予想される。

以上のように，日本語のコーパス言語学では，国語研究所の語彙調査を源流とする理論モデル主導型のアプローチが，BCCWJ という大規模均衡コーパスとして結実する一方，データ主導型の研究を行う環境も整いつつあるように思われる。ただし，両者の違いは，それぞれの研究成果の蓄積がいまだ多くないこともあって，研究者の間でも明確に認識・自覚されておらず，十分な検討が行われるに至っていない。

６．探索的コーパス言語学の方法

以上の検討から，筆者は，コーパス言語学においても，統計学およびそれを科学的探究の方法として採用する多くの応用分野と同様に，理論モデル主導型からデータ主導型への展開が生じている，日本語のコーパス言語学に限れば，生じつつある，と考える。このことは，コーパス言語学，とりわけ日本語のコーパス言語学にとって，重要な意味をもつだろう。なぜなら，BCCWJ のような大規模均衡コーパスによって理論モデル主導型の研究が推進される一方で，データによってデータそのものを説明する（データ自身に語らせる）ようなデータ主導型の研究が，今後大きく展開していく可能性が十分にあるからである。

このような現状において，日本語のコーパス言語学に求められるのは，すでに方法論として確立している大規模コーパスによる理論モデル主導型の研究とは別に,なお開発途上にあるデータ主導型の方法論を追究していくことだろう。ただし，ここには大きな問題がある。それは,「データによってデータそのものを説明する（データ自身に語らせる)」とは具体的にどのようなことであり，また，そのためにはどのような方法を用意すればよいのかという問いに対して，一義的な答えは用意できないということである。渡部ほか(1985：4)によれば,「探索的データ解析の筋道は明確に定まっているものではなく，任意的であ」るという。このことは，おそらく，基本的にデータ主導型のアプローチ全体にあてはまるものと考えられる。欧米のコーパス駆動型言語学の内容が必ずしも

定まらないのも，こうした側面が関係しているのかもしれない。そして，おそらく，この問題に対する答えは，分野によって，また，研究者によっても異なるのだろう。

荻野綱男は,「計量言語学の観点から語彙研究を展望し，今後の語彙研究で必要となる点を指摘する」論文の中で,

> 多人数調査による計量的な語彙研究の面では，簡単な手法を用いたものや，数量化の諸手法など「データから何かを探ろうとする」アプローチが多く，仮説検証的な実験法によるアプローチがほとんどない。今後は，この方面の研究が多くなることが必要である。（荻野 2002：97）

とし，研究開始前の事前の仮説がなく，調査をした上でそれを探ろうとする「調査」スタイルの研究は，研究仮説と（実験計画による）その検証という形で目的や手段がはっきりと意識されている「実験」スタイルの研究に比べて,「中途半端」で「場当たり的」であり,「多くの研究が成果の積み重ねにならず，そのたびに新しい地平で開始されなければならなかったのではないか」と述べるが（荻野 2002：105-106），これも，探索的なアプローチの方法論における任意性を批判したものだろう。

であればこそ，データ主導型アプローチにもとづく「探索的コーパス言語学」の方法論を開発していくことの必要性は大きいものと考える。こうした問題を認めた上で，あくまでも一つの試みに過ぎないが，本稿では，以下の 4 点をその方法論として提案したい。

(1) 連結パターンの探索
(2) 全文コーパスの利用
(3) 多様なコーパスの開発
(4) 探索的データ解析の適用

このうち，(1)と(2)は「内的方法」とでも呼ぶべきものである。(1)の「連結パターン」(association pattern) とは，バイバーほか(2003)による用語で,「言語項目（linguistic feature）が他の言語項目や非言語項目と連結して用いられる方法」をさし，問題の言語項目が文脈の中で特定の単語と連結されたり（語彙的連結），同じく特定の文法構造と連結されたり（文法的連結）する「言語的連結」と，問題の言語項目がレジスター・方言・史的期間といった非言語

的側面と関連し特有の分布をする「非言語的連結」とに分けられる。バイバーほか(2003：10)は,「コーパスにもとづく分析の本質的な特徴とは，言語項目の用法を，それと関連のある連結パターンを考慮して研究すること」であるとしているが，この連結パターンこそ，データの探索によって探り出される「データの構造」の一つではないかと考えられる。

(2)の「全文コーパス」とは，対象とする個々の文書のテキストを全文まるごと収めたコーパスであり，連結パターンの探索において，テキストの断片を標本として集めた「標本コーパス」よりも有利であると考えられるものである。もちろん，語彙的あるいは文法的連結を探る場合は，KWIC コンコーダンスを用いたコロケーション分析のように狭い範囲を見るだけで足りることがあり，標本コーパスでも問題ないが，求める連結パターンが連文やより大きな文章（談話）のレベルにまで及ぶと，一つのテキストをまるごと収めた全文コーパスの方が，データに対する洞察をより深めることができるものと考えられる[6)]。

一方，上の(3)と(4)は「外的方法」とでも呼ぶべきもので，(3)は，全文コーパス以外にも様々なコーパスの可能性を考えることによって，探索的コーパス言語学の可能性――コーパスを使って可能になるデータ主導型研究の領域――を広げていく試みであり，(4)は，データ主導型の統計のうち，すでに紹介した探索的データ解析（EDA）の手法をとりあげ，それらが可能にする探索的な日本語研究について検討する試みである。このうち,探索的データ解析については，データ主導型統計学の代表でありながら，これまでコーパス言語学でとりあげられることはほとんどなかった。そのことは,「統計の入門書であると同時に，コーパスを用いた計量的言語研究そのものの入門書でもある」石川ほか編(2010)が，推測統計学と多変量解析に多くのページを割きながら，探索的データ解析については触れるところがないことや，マケナリー&ハーディー(2014：79)が「特定の結果の有意差検定を行う分析ではなく，データ内の構造や関係を調査する……探索的手法」が多数存在すると述べつつも,「コーパス言語学に関して最も重要なものは，因子分析とクラスター分析である」として，探索的データ解析に一切言及していないことなどにも表れている。

以上(1)～(4)の方法の妥当性・有効性に関しては，筆者自身，これまで，いくつかの実践的な検討を行ってきた。内的方法(1)(2)については，主に教科書

と新聞の簡易全文コーパスを用いて，低頻度語（石井 1996・2002），特徴語（石井 2015），臨時一語化（石井 2013a），一語化・語彙化（石井 2012），借用語・訳語（石井 2009）といった語彙レベルの言語項目と文章の構造や展開，話題といった文章レベルの特徴との連結パターンの探索を試みた。また，外的方法のうち，(3)については，いくつかのコーパス（マルチレベル通時コーパス，単一言語パラレルコーパス，マルチメディア・コーパスなど）を試作して，それらが新たにどのような日本語研究を可能にするかを実際の分析事例をもとに検討し（石井 2014），(4)については，探索的データ解析の手法のうち，とくにリジット解析（石井 2016）と蛇行箱型図（石井 2013b）に注目し，これも具体的な分析事例にもとづいて，それらが探索的な日本語研究に有用であることを検討・確認した。

もちろん，探索的コーパス言語学の方法はこれだけではないだろうが，当面は，これらによってどのような日本語研究が可能となり，また，どのような知見が得られるかについての実践的な検討を重ねることで，日本語のコーパス言語学に，大規模標本コーパスにもとづく確認的な（理論モデル主導型の）アプローチだけでなく，多様なコーパスを利用する探索的な（データ主導型の）アプローチもまた必要であることを確認していきたいと考える。

注

1）広義の“corpus-based”は，「コーパス準拠型」と訳され，「コーパスデータやコーパス研究手法を利用する言語研究手法の総称」とされる（マケナリー＆ハーディー2014：357）。

2）この文の原文は，“The corpus, therefore, is seen as more than a repository of examples to back pre-existing theories or a probabilistic extension to an already well defined system.”であり，“a probabilistic extension”は「(確立された既存の理論体系を）拡張しうる手段」とのみ訳されているが，これを文字通り「確率的（に）拡張（するもの)」と解釈すれば，それは（標本から母集団への）統計的推測を意味しているようにも思われる。

3）データ主導型のアプローチは，仮説形成と密接にかかわるものの，仮説形

成そのものではない。仮説形成の思考法についてはＣ・Ｓ・パースの「アブダクション」が知られているが，米盛(2007：53-54)によれば，それは次のようなものである。

> 探究という科学的行為はある問題状況に直面し何らかの疑念を抱くようになるとき，その疑念に刺激されて生じます。つまり探究はわれわれの信念にそむくある意外な事実を観察したり，あるいはわれわれの期待の習慣に反する何らかの変則性に気づくことからはじまります。そして探究の目的は，その意外な事実や変則性がなぜ起こったかということについて，その理由または説明を与えることによって，われわれの疑念を合理的に解決することにあります。このように，ある意外な事実や変則性の観察から出発して，その事実や変則性がなぜ起こったかについて説明を与える「説明仮説」（explanatory hypothesis）を形成する思惟または推論が，アブダクションです。このアブダクションの推論の形式を，パースはつぎのように定式化しています。
>
> 驚くべき事実Ｃが観察される，
>
> しかしもしＨが真であれば，Ｃは当然の事柄であろう，
>
> よって，Ｈが真であると考えるべき理由がある。
>
> ここで「驚くべき事実Ｃ」というのはわれわれの疑念と探究を引き起こすある意外な事実または変則性のことであり，「Ｈ」はその「驚くべき事実Ｃ」を説明するために考えられた「説明仮説」です。（米盛 2007：53-54）

データ主導型のアプローチが，このこのアブダクションという推論形式の中に位置づけられるとすれば，それはその出発点である「ある意外な事実や変則性の観察」（パースの定式化では「驚くべき事実の観察」）にとどまり，「説明仮説」の形成にまで及ぶものではないと考えられる。ただし，科学哲学の世界では，こうした「事実の発見」は「偶然」によることが多く，法則や理論の発見のように「目的志向的な自覚的探究の結果」ではないとされるが（伊東 1981：244），データ主導型のアプローチにおける「探索」とは，どのようなデータであっても，そこにどのような構造が隠されているかを，データそれ自身に語らせるように探り出す行為であり，それは目的志向的・自

覚的な「事実発見」の作業と言えるだろう。

4）国語研究所の「新聞三紙の語彙調査」では，調査期間中から『電子計算機による国語研究』という報告書がシリーズで刊行され，調査に従事した研究者による数多くの論考が発表された。語彙調査としての「失敗」の経験の中で生み出されたともいえるこれらの論考が，その後の日本語研究（日本語学，計量言語学，コーパス言語学など）の発展に大きく寄与し，また，機械翻訳や人工知能などの言語処理研究にも影響を与えたことは，強調しておきたい。

5）http://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html，2017 年 10 月 3 日アクセス。

6）マケナリー＆ハーディー(2014：227-228)は，

> 一般的に言って，コーパス言語学の基盤となっている分析技術は，実際のところ，コンコーダンス分析やコロケーション分析のように極端に狭い範囲を対象とするか，あるいは，頻度リスト作成やキーワード検索のように，テキストデータから離れ，談話の連続性を無視して，異なるテキストから得られた結果を統合する抽象的なものになるかいずれかである。前者の場合，検索対象とした箇所をせいぜい数十語のコンテキスト内で調べるだけで，それ以上のコンテキストを見ることはほとんどないのだから，標本を完全テキストにする必要性を主張するのは難しい。後者の場合も，完全テキストからなる標本群が，冒頭部・中間部・末尾部等から抜粋した標本を均衡的に集めたものに勝る理由を明らかにするのは難しい。

とするが，この見方は，コーパスを用いた言語研究の対象領域をあまりにも狭く限定しているように思われる。

引用文献

石井正彦(1996)「使用頻度“１”の語と文章－高校『物理』教科書を例に－」『国立国語研究所研究報告集』17:23-55，秀英出版

石井正彦(2002)「第９章　語彙と文章」『朝倉日本語講座４　語彙・意味』191-207，朝倉書店

石井正彦(2009)「借用の位相―ＪＳＴ・科学技術文献情報の「ユビキタス」を

例に―」『待兼山論叢 日本学篇』43:73-90
石井正彦(2012)「『不良債権処理』ができるまで―新聞にみる語彙化現象の動態―」『日本語学』31(14):50-61,
石井正彦(2013a)「臨時的な四字漢語の形成―文章論的な視点から―」野村雅昭編『現代日本漢語の探究』146-166, 東京堂出版
石井正彦(2013b)「探索的データ解析による言語変化研究―蛇行箱型図によるＳ字カーブの発見―」相澤正夫編『現代日本語の動態研究』129-150, おうふう
石井正彦(2014)「第 4 章　多様なコーパスの可能性」『講座日本語コーパス 6 コーパスと日本語学』69-101, 朝倉書店
石井正彦(2015)「無性格語は実在するか―特化係数とその散布度による検討―」斎藤倫明・石井正彦編『日本語語彙へのアプローチ－形態・統語・計量・歴史・対照－』147-163, おうふう
石井正彦(2016)「リジット解析－計数データを用いた言語研究への適用－」『計量国語学』30(6):357-377
石川慎一郎・前田忠彦・山崎誠編(2010)『言語研究のための統計入門』くろしお出版
伊東俊太郎(1981)「科学的発見の論理」『科学と現実』239-261, 中央公論社
荻野綱男(2002)「計量言語学の観点から見た語彙研究」『国語学』208:97-115
田野村忠温(2000)「用例に基づく日本語研究－コーパス言語学－」『日本語学』19(5):192-201
田野村忠温(2014)「BCCWJ の資料的特性－コーパス理解の重要性－」『講座日本語コーパス６　コーパスと日本語学』119-151, 朝倉書店
土屋信一(1989)「語彙調査－全体的な見通しとねらい－」国立国語研究所『高校中学校教科書の語彙調査　分析編』1-5, 秀英出版
冨田　隆・三輪和久(2002)「発見における有効な仮説検証方略と協同の効果」『認知科学』9(4):501-515
Ｄ・バイバー, Ｓ・コンラッド, Ｒ・レッペン (2003)『コーパス言語学―言語構造と用法の研究―』(齊藤俊雄ほか訳) 南雲堂 (Biber, D., Conrad, S. and Reppen, R (1998) *Corpus Linguistics: Investigating Language*

Structure and Use. Cambridge University Press. ）

Ｔ・マケナリー，Ａ・ハーディー（2014）『概説コーパス言語学－手法・理論・実践』（石川慎一郎訳）ひつじ書房（McEnery, T. and Hardie, A. (2012) *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.）

宮島達夫(2007)「語彙調査からコーパスへ」『日本語科学』22:29-46

吉田忠編(1995)『現代統計学を学ぶ人のために』世界思想社

米盛裕二(2007)『アブダクション　仮説と発見の論理』勁草書房

渡部洋・鈴木規夫・山田文康・大塚雄作(1985)『探索的データ解析入門－データの構造を探る－』朝倉書店

Gilquin, G. and Gries, St. Th. (2009) Corpora and experimental methods: a state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5(1), 1-26.

Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

（文学研究科教授）