語彙・計量研究

『日本語科学』24(2008年10月) 117-124　〔小特集〕国立国語研究所の60年　[寄稿論文]

語彙・計量研究

石井　正彦
(大阪大学)

1. 言語使用研究としての語彙調査

ここで「語彙・計量研究」とは，国語研究所が行ってきた語彙調査に代表される計量的な語彙研究を主としてさすものとし，それに，近年の大規模なコーパス（の構築）にもとづく語彙研究を含めるものとしよう。「語彙研究」とは，本来，計量的なそれに限らないし，また，「計量研究」も，広く統計的・数理的な言語研究と考えれば，語彙研究に限られるものでもない。実際，国語研究所には，そうした語彙研究・計量研究が数多くある。しかし，今回の小特集の部立てをみるかぎり，方言研究，社会言語学，国語教育・政策，日本語教育に並ぶものとしての「語彙・計量研究」とは，たとえば文字・表記や文法の研究に対する単語や語彙の研究という意味ではなく，国語研究所で最も大規模かつ組織的に行われた語彙調査とコーパスの構築，そして，それにもとづく計量的な語彙研究をさすものと解すべきだろう。

さて，言語の研究を，大きく，その体系・構造の側面に注目するものと，使用の側面に注目するものとに分けるとすれば，語彙調査は，調査対象とする文章・談話の集合でどのような単語がどれほど使われているかを観測するものだから，基本的には，言語使用（語彙使用）の研究に位置づけられるはずである。しかし，一方で，語彙調査の結果が，基本語彙や分類語彙表など，語彙の体系・構造面の組み上げに利用されることも普通であり，むしろ，語彙調査を言語使用の研究と言い切ることの方が一般的ではないかもしれない。確かに，語彙調査というと，（単語の語彙的性質や語彙の体系的性質を明らかにする）語彙論の道具立てとみなされることが多いが，そこには，言語使用の研究として，語彙論以外の他の分野とも重なる側面を見出すことができるのも事実である。以下では，まず，これまでの国語研究所の語彙調査に，そうした言語使用研究としての側面が見出せることを確認していこう。

2. 社会言語学的な側面

語彙調査が言語使用の研究としてあるならば，それは，当然，言語使用を規定する社会的側面の研究，すなわち，社会言語学にかかわることになる。創設当初に白河市や鶴岡市で行われた「個人の一日の言語生活」における使用語彙の調査（報告2・5）は，どのような社会的条件をもつ話者が，一日の言語生活で，どのような単語を，どれほど使っているかを調査したもので，まさに，社会言語学的な語彙調査といえる。その後も，知識階層（日本語教育・語学関係の研究者）を対象とした調査（野元菊雄ほか 1980）で，発話場面（公的生活・私的生活・外出先）と使用語彙との関係が分析されたり，テレビ放送の語彙調査（報告112）で，番組特性と並んで，

話者の性別・年齢・職業などと使用語彙との関係が追究されたりしたのも，社会言語学的な調査といってよいだろう。なお，談話語の実態調査（報告８）では，日常談話を中心とした語彙調査も行われているが，それは話しことば特有の語彙使用を見出すためのもので，社会言語学的な側面はない。

書きことばの語彙調査に社会言語学的な側面を見出すことは難しいが，雑誌や新聞の語彙調査（報告 21・37）は，テレビとあわせて，マス・コミュニケーションの言語を対象とするものであり，それ自体，社会言語学的であるといえる。ただし，これらは，調査年代がだいぶ離れているので，相互の比較は行われていない。ほかに，知識体系の記述語彙を調べた高校・中学校教科書の語彙調査（報告 76・87）も，基礎的とはいえ専門世界の言語を扱ったものとして社会言語学的である。

3. テキスト言語学的な側面

単語（語彙）は，ひとまとまりの文章・談話を構成するために使われるのであるから，語彙調査が，言語使用の，文章・談話のあり方に規定される側面，すなわち，テキスト言語学的な側面を問題とすることは当然である。この問題は，広く，書きことばと話しことば，さらに，書きことばにおける雑誌・新聞・教科書など，異なるテキストタイプ間の言語使用の違いと考えることもできるから，国語研究所の語彙調査を横断的に眺めれば，それぞれに特徴的な語彙使用を見出すことも可能である。石綿敏雄（1989）は，そうした観点から，雑誌・新聞・教科書における語彙使用の共通面と相違面とを指摘したものであり，林四郎（1982）の「臨時一語」も，新聞文章の大量生産的な特徴と臨時的な語形成とのかかわりを論じたものである。なお，書きことばと話しことばとの比較は，社会言語学でいう「レジスター（言語使用域）」としての言語使用の変異を調べているとみなすこともできる。

また，この種の問題は，それぞれの語彙調査の中で設定した「層別」の使用語彙を比較することによっても検討できる。たとえば，雑誌 90 種の調査（報告 21）では，対象とした雑誌を５層（評論・芸文，庶民，実用・通俗科学，生活・婦人，娯楽・趣味）に区分し，新聞３紙の調査（報告 37）では，「文章の種類」によって 17 層，「話題」によって 12 層，「署名態度」によって 10 層，「紙面上の位置」によって８層といった重層的な層別がなされており，それぞれの層に特徴的な語彙使用を見出すことができる。こうした層別は，国語研究所のほとんどの語彙調査で行われており，その意味で，語彙調査におけるテキスト言語学的な側面の表れといえる。

しかし，語彙使用における，より厳密な意味でのテキスト言語学的な側面とは，単語が，ひとまとまりの文章・談話の構成や展開にどのように使われているのか，その具体的な様相を明らかにすることであろう。国語研究所の語彙調査の中で，こうした側面の追究を明確に掲げたのは，そのために全数調査を採用した，教科書の語彙調査（報告 76・87）のみである。物理の教科書を使って，文章における話題の展開と語彙使用の変化との関係を追及した中野洋（1980）や，同じく，低頻度語の出現を規定する文章上の諸特徴を探った石井正彦（1996）などは，この語彙調査におけるそうした試みの一つである。

4. 歴史言語学的な側面

語彙の使用は，また，それが使用された時代にも規定されているから，語彙調査は，たとえある一時点の調査であっても，歴史言語学的な側面をもつことになる。これは，「郵便報知新聞」の語彙調査（報告15）や国定読本の調査（コンコーダンス作成）（国語辞典編集資料1～12）などでは，近代語から現代語への変化ないし現代語の源流を探るものとして，当初から付与された側面であるが，現代語の語彙調査であっても，それがいつの時点での語彙使用を調査したものであるかは，つねにその結果を規定することになる。

語彙調査がこうした歴史言語学的な側面をもつことから，ある時点での語彙調査に対して，後年，共通の対象を同じ方法で調査することにより，二つの時点間の語彙使用の，歴史的な変化を見出すことも可能になる。1994年の月刊雑誌70誌の調査（報告121）は，その時点での雑誌の語彙使用をみるとともに，1956年の雑誌90種の調査（報告21）と比較して，ほぼ40年の間の語彙使用の変化を見出すことも目標としている。ただ，この場合も，40年近い年月を経て，共通の調査対象（雑誌）をとることは困難でもあり，異なる語彙調査の完全なる比較は容易ではない。

今のところ，一つの語彙調査で通時的な調査を行ったものは，1906年から1986年まで10年おきに各年1万語ずつを標本とした『中央公論』の調査（報告89・石井久雄1990）しかない。これは，調査対象を同一の雑誌としたほか，各年代を同じ調査単位・方法で調査することにより，語彙使用の違いを（最大限）時代的な変化としてとらえられるようにしたものである。

また，語彙調査は行われていないが，雑誌『太陽』のコーパス（資料集15）も，1895年から1925年までほぼ8年刻みで各年300万字程度の記事を収めており，「確立期日本語」の言語変化を追究することができる。

5. 語彙還元論的な見方

以上のように，国語研究所の語彙調査には，言語（語彙）使用の研究として，社会言語学，テキスト言語学，歴史言語学に重なる側面を見出すことができる。しかし，一方で，語彙調査には，語彙の使用と社会・テキスト・歴史との関係を捨象して，語彙使用は語彙そのものの本質的な特徴によって決められるとする，いわば語彙還元論的な見方がある。

国語研究所の，とくに書きことばの語彙調査では，多くの場合，どんな単語がどれほど用いられているか（使用率と使用範囲）を調査し，それをもとに単語の基本度（基本性）を導いて，基本語彙を選定する（ための基礎資料とする）ということが目標とされている。では，なぜ，単語の使用率や使用範囲を調べることが，その基本度を測ることになるのか。それは，単語の使用率・使用範囲を，その単語自身の基本度が反映した現象であると考えるからである。つまり，より基本的な単語はより多く・広く用いられ，あまり基本的でない単語はより少なく・狭く用いられる。その単語がどのように用いられる（現象する）かは，その単語の本質としての基本度が決定すると考えるのである。これは，語彙の使用を語彙の本質によって説明する，語彙の自律的・還元論的な側面を重視する見方であるといえる。

しかし，こうした見方は，言語使用を社会・テキスト・歴史などから切り離し，言語のみによって説明しようとするもので，問題がある。語彙教育の世界では，学習者が接触する機会の多い文章·談話で高使用率·広範囲に使われる単語は，教育上，重要な単語である。しかし，それは，優先的に指導ないし学習（習得）した方がよいという意味で重要だということであって，それだけで，その単語の本質的な特徴の重要性を確認するものではない。教育基本語彙を使用率・使用範囲をもとに選定しようとするなら，そこには教育上の実利的な重要性が反映されているのであり，単語そのものの重要度が表れているわけではない（したがって，阪本一郎（1965）では，この方法による選定を採用していない)。

国語研究所の語彙調査でも，得られるのは，調査対象とした文章・談話で使用＝接触確率の大きい単語（語彙）にすぎないはずである。なぜ使用＝接触確率が大きいかといえば，それは，その単語が本質的に基本的であるからではなく，その単語を繰り返して使用することを，社会，テキスト，歴史にかかわる諸側面（のいずれか）が後押ししたからである。語彙使用の研究にあっては，そうした後押しの様子を具体的に明らかにしていくことが必要だろう。強いていうなら，基本語彙ではなく，基本的な語彙使用をこそ，追究すべきである。

なお，単語の基本度，および，それにもとづく基本語彙という考え方は，基本度というものを使用と切り離して説明しなければ，循環論になる。そのせいもあって，語彙調査では，単語の基本度には，使用率・使用範囲だけではなく，語彙体系のどこに位置するかということも関係するとされた。婦人雑誌（報告４）・総合雑誌（報告 12）・雑誌 90 種の調査（報告 21）では，この面を明らかにするために，使用率順語彙表・五十音順語彙表とともに，分類語彙表がつくられた。ただし，語彙調査で得られた高使用率・広範囲の語彙だけで分類語彙表の意味分野全体を覆うことはできず，後に別に編まれた『分類語彙表』では，阪本一郎（1965）から多くの単語を補っている。

6. 抽象的な単位の計量

単語（語彙）は，文章・談話の中で，偶然に，あるいは，ランダムに使われるのではない。それは，上に見た，社会，テキスト，歴史などにかかわる諸側面に規定されながら，必要に応じて，繰り返して使われたり，使われなかったりする。つまり，語彙使用には一定のパターンがあるのであり，それは，繰り返して使われる（使われない）という量的な傾向となって，社会，テキスト，歴史などにかかわる諸側面と語彙使用との関係を表現している。こうした量的傾向を観察するために，語彙調査は，計量という操作を行う。しかし，語彙調査の計量は，基本的に語彙還元論的な見方に立つもので，語彙使用のパターン＝量的傾向を十全にはとらえることができない。

語彙調査では，まず，文章や談話（の断片）がそれを構成する単位語に切り分けられ（単位切り），それらがしかるべき基準のもとに見出し語にまとめられて（同語異語判別），一つの見出し語がいくつの単位語をもつかということが（見出し語の使用頻度として）カウントされる。このとき，文章・談話を構成する単位語は，二重に言語使用から切り離される。すなわち，最初に，

それが含まれる文脈から切り離されて，前後の語とともにつくりあげていた文脈的意味を失い，次いで，抽象的な単位としての見出し語にまとめられることによって，文脈的意味に関与していた語彙的意味（の区別）をも失うのである。「頭が割れるようにいたい」「もっと頭を使え」「鼻の頭に汗をかいた」「頭を刈ってもらった」から見出し語「頭」をとりだして，その頻度をカウントしても，それは，「頭」という単語がなんらか（いずれか）の意味で使われた回数を表すだけで，どのような意味で何回使われたかを示すことはできない（これに対して「セマンティック・カウント」も提案されているが，行われていない）。結局，語彙調査における見出し語の意味は，それにまとめられたすべての単位語の文脈的意味・語彙的意味を抽象したものとしか言いようがなく，それが個々の使用で（他の単語とともに）どのような意味を表しているかというパターンは明らかにされない。

こうした語彙調査の計量は，具体的な語彙使用を抽象的な見出し語（の頻度）に還元してしまうやり方であり，語彙還元論的な見方に立つものといえる。もちろん，語彙調査でも，KWICなどを利用してコロケーションの分析などを行うこともできるが，それは語彙調査の本来的な目標ではない。

7. 確認と探索

語彙調査が，語彙使用の具体面を捨象した抽象的な単位としての見出し語を計量するという点は，標本調査としての語彙調査が，基本的に，標本における語彙使用から母集団におけるそれを推定するという考え方に立っていることとも対応する。語彙調査において本当に知りたいのは母集団の様子であり，目の前にある標本はそれを知るための手がかりにすぎない。そして，そのようにして推定される母集団の語彙使用とは，やはり，具体的な語彙使用ではなく，抽象的な見出し語としてのそれであろう。このような，標本から母集団を推定する統計は，推測統計学（推計学）と呼ばれ，語彙調査では，雑誌90種の調査が最高の水準に達しているといわれる。そこでは，「使用率の小さい見出し語については標本使用度数の変動が大きく，標本に現われたか否かが相当に偶然に左右されるという理由」（報告21，p.21）により，標本使用度数が7以上の見出し語しか載せない語彙表がつくられた。

アメリカの統計学者テューキーは，こうした統計的推測を中心とする立場を「確認的データ解析」と呼ぶ一方で，標本と母集団という関係をひとまず措き，データそれ自体を独立した情報源と見て，そこから最大限の情報を引き出し，データに潜む問題点や特徴を探索していこうとする立場を「探索的データ解析」と呼んだ。テューキー自身のたとえによると，確認的データ解析は，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り，探索的データ解析は，証拠を収集して立件していく刑事や警官のような仕事であるという（吉田忠編1995:104）。高頻度語の使用だけでなく，低頻度語の使用についてもその特徴を見出そうとする探索的データ解析は，言語使用の具体面を探る有効な統計手法ではないかと考えられる。なお，語彙調査では，テレビ放送の調査（報告112）が探索的データ解析の手法を部分的に採用しているが，必ずしも十分とはいえない。

8. 言語（語彙）使用のパターン

6節で触れた，見出し語の意味が具体的に特定できないという問題は，単位語を見出し語にまとめることにもよるが，より基本的には，上述したように，文章や談話（の断片）をすべて単語（ないし形態素）に切り分けてしまうことによるものである。単語は，確かにそれ自身で意味をもつが，文章·談話の文脈の中では，他の単語と結びついたより長い単位＝句(コロケーション)の中にあって，新たな意味をつくりだしていることが多い。そして，そうした結びつきとその意味も，また，まったく偶然につくられるのではなく，われわれにとってある程度予測可能な，一定のパターンをなしていることが多い（Stubbs 2002)。

たとえば，いま，『ＣＤ－毎日新聞データ集　2002年版』で，「頭（あたま)」という語を「～が」という形式で検索してみると，「頭がある」「頭がいい／よい」「頭がいっぱい」「頭がすっきりする」「頭が下がる」「頭が固い」「頭が重い」「頭が上がらない」「頭が真っ白（になる)」「頭が痛い」「頭が白くなる」など，ごく限られたパターンでしか使われていないことがわかる。

また，たとえば，「人々」という単語が，『同データ集　2000年版』で「～の人々」という形式で用いられたとき，「～」の部分にどのような単語が使われているかを調べると，①「アジア」とくに「台湾」「中国」「韓国」「北朝鮮」など東アジア諸国が多く，「アメリカ（米国)」は少ない，②「東ティモール」「チベット」など紛争のある国が多い，③「(開発）途上国」が多く，「先進国」は少ない，④日本国内では，「沖縄」が圧倒的に多く，「神戸」「長崎」も見られるが，「東京」は１例もない，⑤「村」や「町」が多いが，「都会」は少ない，⑥「一般」「普通」「無名」などが多い，⑦「ホームレス」「在日」などが多い，⑧「世界」「全世界」「世界中」などが多い，といったことがわかる。つまり，「人々」という単語は，新聞記事というテキストの中で，書き手から見て，過去に侵略して「申し訳ない」とか，紛争・貧困・被災・差別などの理由で「気の毒だ」とか，普通で名もないとか思う対象に向けて，そうした暗示的意味のもとに使われることが多いのである（石井正彦 2004)。

このように，特定のテキストの中で，単語が他の単語とつくるパターン化した結びつきや，そこでつくられる暗示的意味は，文章や談話をすべて単語に切り分け，見出し語に抽象するという語彙調査では，見出すことができない（見出すことを第一の目的とはしていない)。こうした，コロケーションやテキストに依存した語彙使用の側面は，コーパスを用いた言語使用の研究によってなされるものだろう。単語（語彙）がいかに使用されているかを，より具体的・詳細に記述するためには，その使用を文章・談話の中でそのまま，まるごととらえる必要があり，そのためには，コーパスと，それにもとづく語彙使用の研究が必要なのである。

9. 語彙調査からコーパスによる言語使用研究へ

Stubbs (2002）は，コーパス言語学は，本質的に，社会言語学的であり，通時的であり，計量的であるとしている（邦訳:309)。これらの側面は，すでに見たように，テキスト言語学的な側面も含めて，国語研究所の語彙調査にも認められるものであり，その意味では，語彙調査とコーパス言語学との間に大きな違いはない。しかし，語彙調査の背景には，これもすでに述べたよう

に，語彙還元論的な見方があり，語彙使用の具体面が少なからず捨象されてしまうという問題がある。コーパスを用いた言語使用研究は，語彙調査のそうした問題点を克服しつつ，上の諸側面を前面に押し出しながら，言語使用をより具体的に追究するものとして展開されるだろう。そして，さらには，コーパスにおける言語使用を，言語の本質が現象した「用例」としてではなく，人間が社会的な相互作用の中でつくりあげる「言説」ととらえ，そうした言説の中に人間がどのような意味をつくりあげているかを探る，構築主義的な言語使用研究にまで発展していく可能性をも感じるのである。

国語研究所では，太陽コーパス（資料集 15），日本語話し言葉コーパス（報告 124）に続いて，現在，書きことばの大規模均衡コーパス（シンポジウム報告 13）が作成されつつある。これらを用いた言語使用の研究は，語彙調査のそれを大きく超えたものになるだろう。

参考文献

（国立国語研究所の報告書等）

報告 2『言語生活の実態―白河市および附近の農村における―』（1951）
報告 4「現代語の語彙調査　婦人雑誌の用語」（1953）
報告 5『地域社会の言語生活―鶴岡における実態調査―』（1953）
報告 8『談話語の実態』（1955）
報告 12『総合雑誌の用語（前編）―現代語の語彙調査―』（1957）
報告 15『明治初期の新聞の用語』（1959）
報告 21『現代雑誌九十種の用語用字（第 1 分冊）総記および語彙表』（1962）
報告 37『電子計算機による新聞の語彙調査』（1970）
報告 76『高校教科書の語彙調査』（1983）
報告 87『中学校教科書の語彙調査』（1986）
報告 89『雑誌用語の変遷』（1987）
報告 99『高校・中学校教科書の語彙調査　分析編』（1989）
報告 112『テレビ放送の語彙調査 I』（1995）
国語辞典編集資料 1 ～12『国定読本用語総覧 1 ～12』（1985-1997）
報告 121『現代雑誌の語彙調査―1994 年発行 70 誌―』（2005）
報告 124『日本語話し言葉コーパスの構築法』（2006）
資料集 15『太陽コーパス　雑誌『太陽』日本語データベース』（2005）
シンポジウム報告 13『言語コーパスの構築と活用』（2006）

（その他）

石井久雄（1990）「『中央公論』1986 年の用語」『国立国語研究所研究報告集-11-』，1-40，国立国語研究所
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，国立国語研究所
石井正彦（2004）「コーパス言語学と『キーワード』」『月刊言語』33（12），90-91，大修館書店
石綿敏雄（1989）「雑誌・新聞語彙と教科書語彙」『高校・中学校教科書の語彙調査　分析編』，

6-14，国立国語研究所
阪本一郎（1965）『教育基本語彙』牧書店
中野洋（1980）「文章における語彙の構造に関する探索的研究（4）―初出語の分布―」国語研究所内部資料『季報』1980 春号
野元菊雄ほか(1980)『日本人の知識階層における話しことばの実態』文部省科学研究費特定研究「言語」研究報告書
林四郎（1982）「臨時一語の構造」『国語学』131，15-26，国語学会
吉田忠編（1995）『現代統計学を学ぶ人のために』世界思想社
Stubbs, Michael（2002）*Words and Phrases: Corpus Studies of Lexical Semantics*, Oxford: Blackwell.(邦訳『コーパス語彙意味論　語から句へ』研究社)