数理的研究

西端幸雄

1. はじめに

今期（1998—1999）の「数理的研究」を展望するにあたり，その調査範囲は，国語学・日本語学系の雑誌，紀要，報告書に重点を置き，『自然言語処理』『情報処理学会論文誌』などといった理工系の色彩の強い書籍，雑誌，紀要などにはあまり重点を置かなかった点をお断りしておく。

まず，今期の「数理的研究」の動向を知る手がかりを『計量国語学』に求めてみると，1995年6月（20-1）から1998年3月（21-4）までの間に掲載された論文，調査報告，紹介，ノート，随想は，延べ38編であり，そのうち，何らかの統計的な解析を行った，数理的研究の範疇に入る論文等は，19点であった。それが，今期1998年6月（21-5）から2000年3月（22-4）までの間に掲載された論文等は，延べ25編で，そのうち，数理的研究の範疇に入る論文等が22点にものぼり，かなり活発に研究が行われてきていると言えよう。

2. 語彙に関わる研究

日本語のシソーラスの代表的なものとしては，『分類語彙表』(1964.3)があるが，その増補版である中野洋『「分類語彙表」形式による語彙分類表』（文部省科学研究費報告書，1997.3)が世に問われ，すでに2年余りが過ぎた。しかし，管見の限りでは，それを直接に利用した論文などがない。それは，モニター制で極めて狭い範囲に配布されたためかもしれないが，その利用は，今後促進されるべきであると考えている。ただ，その増補版を利用するにあたっては，中間報告として出されたものであるため，内部に見出しや漢字・仮名表記の不統一や誤りがかなり残存している点で，十分な注意を払う必要があると考える。また，その増補の推進者中野洋氏を失った今，この増補版の補訂作業を受け継いで推進する陣容をぜひ整えていただきたいものと切望している。

今期，シソーラス的な性格とコーパス的な性格を併せ持った，天野成昭・近藤公久『NTTデータベースシリーズ　日本語の語彙特性・第Ⅰ期・第Ⅱ期』（三省堂，1999.12・2000.3）が刊行された。第Ⅰ期は，『新明解国語辞典　第四版』の見出し項目を基本とす

る8万余語について，単語とその中に現れる文字（JIS X 0208）に関する各種の調査をまとめた8つのコーパスで構成されており，それぞれのコーパスのすべてのデータ，および親密度評定に使用した音声ファイルもふくめてCD-ROMとして公開されている。一方，第II期は，朝日新聞の1985年から1998年の14年分の記事データに対し，形態素解析を行って単語を切り分け，抽出した36万語の出現頻度と，そこに含まれる文字の出現頻度を調査した結果をまとめたコーパスである。これらのデータと，先に刊行された，日本語の語彙30万語を3000種類の意味属性で分類したシソーラス，池原悟他『日本語語彙大系』（岩波書店，1997.9）とを併せ利用することにより，シソーラスやコーパスを用いた語彙研究・文字研究がより一層活発に，高度になると期待できる。

さて，旧版である『分類語彙表』のデータを用いた研究は，過去にも多く見られるが，今期の範囲では，鈴木泰「『古典対照語い表』のコードづけ――「来」を後項素として含む複合動詞の場合――」（名古屋大学『国際シンポジウム比較語彙研究』，1998.10）や広瀬英史「日本語語彙史研究試論――源氏物語による――」（同上），志甫由紀恵「歌謡曲の語彙――演歌とポップスのいま・むかし――」（大阪樟蔭女子大学『日本語研究センター報告』8，2000.3），金明哲・宮本加奈子「ラフな意味情報に基づいた文章の自動分類」（重点領域研究「人文科学とコンピュータ」『シンポジウム 人文科学における数量的分析（4）』，1999.3）などがあげられる。

このうち，鈴木泰1998.10は，『源氏物語』と『万葉集』に使用されている「～来」という複合動詞の比較を通して，古典語の複合動詞に対してどのように分類コードを付けたらよいのかという試案を提示している。この試みは，筆者も以前行ったことがある。その作業の中で最も判断に窮したのは，複合動詞をコード化するという点であった。例えば，「出で来」の場合，単に後の要素「来」によって機械的にコード化すると，語句全体の表す意味が表出できなくなり，また一方，前の要素「出づ」によってコード化したとしても，「来」という意味合いが薄れてしまう。このように，複合動詞に分類コードを付ける場合，特に後接する動詞が抽象化しているかどうかの厳密な判定を行う必要がある。そうした点の解決の糸口を見つける方策の一つとして，田島毓堂『比較語彙研究序説』（笠間書院，1999.10）が大いに参考になろう。

また，志甫由紀恵2000.3は，1953年から1998年までの演歌とポップスの歌詞をコード化することにより，まず，演歌と若者向けのポップスとの違いを明らかにし，さらに10年単位のそれぞれの使用語彙の性格の変遷を見ようとしたものであるが，両者ともに60年代と80年代に大きな変化が見られる点を述べている。同じように，長期間にわたる歌謡曲の歌詞の語彙を研究したものに鷲岡昭夫「歌謡曲の五十年」（『日本語学』14-9，1995.8）などがあるが，今後も考現学的な観点から，こうした歌謡曲の使用語彙の研究を行うことは価値あるものと考えられる。

語彙を数理的に研究したものとしては，上田英代・村上征勝・藤田真理「源氏物語の

会話文と地の文をめぐる数量分析――助動詞を中心に――」(『計量国語学』21-5，1998.6)，村上征勝・今西祐一郎「源氏物語の文章の計量分析」(重点領域研究「人文科学とコンピュータ」『シンポジウム人文科学における数量的分析(4)』，1999.3)，加藤妙子「大正期童謡の語彙――童謡作品の共通性――」(名古屋大学『国際シンポジウム比較語彙研究』，1998.10)，佐野仁美「外来語の略語の語構成――音における「同化」の分析――」(『国文目白』38，1999.2)，橋本和佳「高校国語教科書〈学習〉欄の外来語」(大阪樟蔭女子大学『日本語研究センター報告』8，2000.3)，佐竹秀雄「新聞第一面の語彙――1997年の新聞3紙を資料として――」(武庫川女子大学『言語文化研究所年報』10，1998.3)，井上博文「ペットの名前の造語法――犬と猫の場合――」(大阪教育大学『学大国文』41，1998.3) など，数多くの論考を目にすることができる。

まず，上田英代他1998.6は，『源氏物語』中の会話文と地の文で使われている助動詞の違いを明らかにしたものである。それによると，会話文においては，「む」「べし」などといった話し手の主観に関わる助動詞が多用され，地の文では，「たり」「り」といった客観性の強い助動詞が多用されているとのことである。こうしたいわゆる付属語は，使用度数が非常に高いため，体系的な研究がなかなか進まないのが現状であろうが，それらを相対的に捉え，数理的に研究することにより，一つの傾向を提示することは非常に有意義なものと思われる。その点で，今後，同時代の他の作品の傾向とも比較することにより，今回提示された傾向が『源氏物語』だけに見られるものなのか，同時代の他の散文仮名文学作品ではいかなる傾向が現れるのかをぜひ示してもらいたいものである。

村上征勝他1999.3は，これまでの継続研究である数量的分析を通して見た『源氏物語』の成立論を，各巻の助動詞の使用状況から補強したものである。

加藤妙子1998.10は，これまで，国語学の研究対象として等閑視されてきた傾向が強い童謡の使用語彙について述べている。特に，北原白秋・野口雨情・西条八十の童謡中に使われているオノマトペに焦点をあて，三者の使用語彙の比較を行っているのであるが，その結果，白秋に際だった特徴が見られると分析している。こうした研究は，その時代の子供に対する見方・考え方が明らかにできるとともに，語彙そのものの位置付けをも明らかにできるものと思われる。その意味で，大正に続く昭和の時代における童謡語彙の特徴，さらには，近現代の童謡語彙の体系的な把握を期待する。

語彙に直接関わらない問題ではあるが，語彙調査を行う際の用例をどの程度調べれば事足りるものなのか，筆者などいつも迷うことがあるが，荻野綱男「用例はいくつ調べればいいか――比率変動推測法の提案――」(『計量国語学』21-5，1998.6)，荻野綱男「用例はいくつ調べればいいか――新出確率推測法の提案――」(『計量国語学』22-1，1999.6) は，そうした迷いを持つ者には大変示唆に富んだ論考である。

荻野綱男1998.6は，用例調査の最初の段階から分類すべきカテゴリーが決定されている場合に適用できるとし，荻野綱男1999.6は，最初の段階では分類すべきカテゴリーが

決定されず，調査の段階的な経過とともにカテゴリーが決定できる場合に適用できるとしている。

3．文字に関わる研究

数理的研究には直接に関わらないとしても，その基礎データとしてのコーパスの構築や語レベル・文字レベルの数理的研究とは絶対的に関わりを持つ文字コードのことについて触れておかざるを得ない。

今期の文字コードに関わる大きな出来事として特記しておくべき点は，いわゆる「新JIS漢字」（第3水準・第4水準：4344文字）JIS X 0213：2000が2000年1月20日に正式に日本工業規格として制定されたことである。1990年に，それまでのJIS X 0208（6879文字）の収録文字数の少なさを克服するためにJIS X 0212（JIS補助漢字：6067文字）が制定され，さらに，それまでのJIS X 0208における文字コードの矛盾を修正したJIS X 0208：1997が世に問われた。今期，JIS X 0213：2000が制定された背景は，コンピュータ利用の多様化の中で，そうした増補と部分修正だけでは対応できなくなったという背景があったからである。

しかし，コンピュータを利用する研究が学際的にも，国際的にも広がりを見せている今日，このようにJIS漢字を増補したとしても，なお不十分であるという感は否めないところである。

そうした点を補完する試みとしては，今期，東京大学を中心にした「GT明朝」（6万余文字），京都大学を中心にした「e漢字」（約12万文字），また文字鏡研究会を中心にした「今昔文字鏡」（約9万文字・バージョンアップ版），TRONプロジェクトを中心にした「超漢字」（約13万文字）などが，データ公開，または，製品版の販売を始めた。

上に掲げた各プロジェクトの一部を紹介したものに，丹羽正之「e漢字」・田村毅「GTコード」（『月刊しにか』，1999.6）があり，「今昔文字鏡」には，今昔文字鏡研究会編『パソコン悠悠漢字術　今昔文字鏡徹底活用』（紀伊國屋書店，1999.4）がある。

このうち，筆者の知る範囲では，「今昔文字鏡」は，大正新脩大蔵経テキストデータベース（略称SAT），インド学仏教学論文データベース（INBUDS），また，筆者が参加している万葉集校本データベース・万葉集関連論文データベースなどの大型プロジェクトにおいて使用されており，その利用度は高い。また，「超漢字」は，TRONプロジェクトの一環として開発されたBTRON-OS（個人ユーザー向けTRON仕様OS）として開発されたもので，ワープロ・表計算・データベースなどの機能を備えており，さらに，前記の「今昔文字鏡」もこの文字体系の中にコード化されている点で注目すべきソフトである。ただ，この「超漢字」自体がOSであるため，当然，基本的にはWindows用ソフトと共用できないという問題点があるが，OSとしては，Windowsに比べ，かなりコンパクトな仕様になっていて，最新の高速・大容量のパソコンでなくても使用に耐える点，今後，

その利用が増える可能性を秘めている。

さて，文字に関わる研究としては，梅田三千雄「部首情報を手掛かりとして確定できる単語の定量的分析」(『計量国語学』21-5，1998.6)，笹原宏之・横山詔一「異体字選択に影響する要因」(『計量国語学』21-7，1998.12)，笹原宏之「漢字字体に対する大学生の接触頻度」(『計量国語学』22-2，1999.9)，宇都宮弘章・都染直也・佐藤栄作『漢字からひらがなへの字形の変化とその数理的解析』(甲南大学総合研究所叢書52，1999.3) などがある。

梅田三千雄 1998.6 は，手書き漢字認識の研究の一部として行われているもので，漢字を構成する偏や旁と，それらとリンクした日本語単語辞書を活用することによって，「偏と旁から構成されている漢字において，その偏が何であるかが認識できると，その単語確定率は約 72.4%になる。同様に，漢字の旁が明らかになると，その時の単語確定率は約 94.6%に達し，極めて高いものとなる」としている。

また，宇都宮弘章他 1999.3 は，毛筆で書かれたいわゆる古筆資料をもとにして，それらのくずしの段階を隷書（漢字）・草仮名・古典仮名（平仮名の意味か）に分けて，隷書レベルの字形データに対して離散型フーリエ変換とベジエ曲線の二つの手法による数理解析を行い，どちらの手法から得られた結果が，実際に手書きされた草仮名や古典仮名に近いかを示したものである。この研究のねらいは，漢字から草体化を経て，平仮名が生成される過程を，これまでの鑑識眼に頼った判定ではなく，システム化することにより，一層説得力を持った説明をする試みの一つである。

ところで，先にも記したが，コンピュータを利用する研究が学際的にも，国際的にも広がりを見せている今日，文字コードの問題に関しては，国内の動向にだけ目を向けているわけにはいかなくなっているのが現状である。その点，我が国だけでなく，西欧やアジアの主要な文字コード，さらに東アジア各国の全漢字コードについて，漢字コード一覧の CD-ROM を付けて分かりやすく解説したものに，安岡孝一・安岡素子『文字コードの世界』(東京電機大学，1999.8) がある。また，平凡社編『電脳文化と漢字のゆくえ』(平凡社，1998.1)では，文字コードが多種多様化することに対する賛否様々な意見が交わされている。

4．その他

今期で，特記しておくべきことは，1995 年度から始まった重点領域研究「人文科学とコンピュータ」が 1999 年度で終了した点である。組織としては，計画班を充実させ，データベース班・テキスト班・イメージ班・数量的分析班の 4 本柱を設けてスタートしたのであるが，当初の目標であった「研究基盤の整備としてデータベースの流通促進とソフトウェアの開発・流通を図り，次に個々の研究の活性化を図るための情報の収集・提供を中心とした研究支援活動の推進を図るということである」という点は，十分に達成されたものと思われる。

その4年間のプロジェクトの中で，数理的研究に関係すると思われる計画班(そのほとんどが数量的分析班）のテーマを以下に掲げておく。

・絵画における色彩美の数量的分析の研究
・文化人類学における家族・親族領域を中心としたフィールドデータの処理と分析
・先史時代ワールドモデル
・西鶴語彙の統計分析のための基礎的研究
・前方後円(方)墳から考察する大和中央政権と東国地方政権との構造的関連
・パーソナルコンピュータによる中期インド・アリアン古文献の解析
・香道伝書の構造と編年の数量的分析の研究
・林の数量化理論第3類による俳句に詠み込まれた音環境の時代変遷の分析
・明治の大新聞・小新聞の語彙変遷に関する数量的分析
・M. ヴィグマンのスケッチの言語的側面に関する数量分析的研究
・「方言認知地図」の計量的処理と分析のためのプログラム開発
・構文解析木を対象とするデータ解析法の研究
・データベース SHUUGE を使用した日本古代の令文解釈諸説の研究――引用形態とその特徴を解析する方法の開発――
・科学読み物の語彙分析にもとづく特徴の抽出
・日本語教育用辞書の適切性に関する数量的解析
・三次元形状計測を用いた考古遺物の形状比較法の研究

今後，このプロジェクトで得られた成果を基に，人文科学系，特に国語学・日本語学の分野でのコンピュータを利用した研究がより一層促進されるものと期待している。

なお，情報処理学会では，2000年10月頃に創立40周年を記念して，創刊号からの40号までの会誌，論文誌（研究会論文誌含む）をすべて収録したCD-ROMを販売する予定である。日進月歩と言われるこの研究領域ではあるが，今日の研究が当然過去の蓄積の上に成り立っていることを考えると，こうした試みは今後の研究にも大きく寄与するものと期待できる。

最後に，筆者は古典の語彙研究の中で，たまたまパソコンを利用している立場でいる。そうした鈍重な生活リズムの中で，この数理的研究という世界を眺めてみると，その研究ペースの速さに目が回るような思いにとらわれることがしばしばであった。そのため，本来必ず取り上げるべき論考を見落としていたり，また，取り上げたとしても，著者の意を汲んでいない点が多々あるかと危惧している。

なお，本稿を成すにあたり，国立国語研究所の江川清氏と同研究所図書館の方々に多大のご援助を賜った。ここに深甚の謝意を表する。

――大阪樟蔭女子大学教授――