数理的研究

荻野綱男

一、はじめに

展望号では、いつも数理的研究の範囲が問題になるが、だいたい毎回の担当者の扱う範囲は一定しているので、筆者もそれにしたがうことにする。

この期には特徴的なことが三点ほどあった。第一に、特定研究「言語の標準化」がスタートしたことである。前回の特定研究(昭52～54年度)のときもそうだったが、今回も、言語を直接の研究対象にする人文系の研究者だけでなく、教育学・医学から工学系にいたる幅広い分野の研究者を含む学際的なプロジェクトになっており、その中には言語の機械処理の研究も含まれている。第二に、新世代コンピュータ技術開発機構(ICOT)が活発な活動を開始したことである。いわゆる第五世代のコンピュータをめざすICOTは、その研究対象の一つとして「言語を理解するコンピュータ」を取り上げており、これに関連して認知科学の一分野としての自然言語理解の研究が非常に盛んになっている。第三に、ワープロが急激に普及したことである(今やディスプレイとプリンタつきで定価が35万円を切るようになった)。ワープロの現状については、『Newton 別冊 ワードプロセッサのすべて最新版』(教育社、昭58・10)にくわしいが、その後の発展もいちじるしい。筆者も技術革新の恩恵にあずかってこの原稿をワープロで打っているが、国語学の分野でも、樺島忠夫氏(『月刊言語』、昭57・2～4月参照)や森岡健二氏などのようにワープロを使いこなす人が出ている。ともかく「言語の機械処理」が一般に行われるようになり、ワープロのための言語研究(たとえば、文法的な解釈を加えたカナ漢字変換などであるが、これも広い意味での数理的研究に含まれる)が主として工学系の人によってなされるようになってきたのである。パソコンの普及・利用も同一線上にある現象であり、言語研究の道具として利用されているが、これについては後述する。

以下では、いくつかの分野別に流れを見ていくことにする。

二、フィールドワークのデータの計量的分析

方言学・社会言語学に関する大規模な調査研究の報告書がいくつか出された。これらから見るかぎり、大規模調査の分析にコンピュータを利用することは、すでに常識になっている。

国立国語研究所(以下国研と略称)は『企業の中の敬語』(三省

堂、昭57・3）と『敬語と敬語意識』（三省堂、昭58・3）を出版した。前者は1086人に対するアンケート調査と254人に対する面接調査の分析である。後者は185人に対するパネル調査と400人に対する継続調査であり、林の数量化理論第三類（以下林3と略称）やAICといった新しい手法も取り込んでいる。井上史雄（編）『《新方言》と《言葉の乱れ》に関する社会言語学的研究』（科研費報告書、自家版、昭58・3）では、10本の論文中2本は理論的なものだが、他の8本はデータの分析結果の報告であり、うち6本はコンピュータを利用している。それぞれの分担者がかなりの規模の調査を行っている。東京大学敬語研究会『続・都市化と敬語』（自家版、昭57・6）は札幌での昭和52年の調査を再分析したものだが、データがコンピュータに入力してあったからこそこのような再分析ができたわけである。荻野綱男「敬語使用から見た聞き手の位置づけの多様性」（『国語学』132、昭58・3）と同「待遇表現の数量化」（朝倉書店『朝倉日本語新講座5運用Ⅰ』、昭58・10）では荻野の数量化の方法を使ってグラフ化までコンピュータで行っている。

福嶋秩子「出雲における開音類の分布とその総合化」（東京大学文学部言語学研究室『言語学演習'83』、昭58・7）同『パソコンによる言語地理学へのアプローチ――SEALユーザーズマニュアル――』（自家版、昭58・2）は、パソコンで地図を作るものだが、地図化はいいとして、そのあと分布の読み取りや解釈をどのように行うのかが大きな課題である。

その答とも関連するが、この期には統計的手法を用いた方言区画論が多かった。井上史雄・河西秀早子「標準語形の地理的分布パターン」（『国語学』131、昭57・12）は因子分析、同「標準語形による方言区画」（『計量国語学』13―6、昭57・9）はクラスター分析、井上史雄「方言イメージ多変量解析による方言区画」（明治書院『現代方言学の課題』、昭58・6）はクラスター分析と林3、井上史雄「共通語的文法表現の地理的分布パターン」（『国語学』133、昭58・6）は林3をそれぞれ利用している。高橋宏一他「青森県言語調査の統計的解析（Ⅰ）」（『Science Reports of the Hirosaki Univ.』29―2、昭57・12）は25の村に対して40問の質問の比率の検定を行い、地域をいくつかに分けている。柴田武「方言区画の方法」（『都立大学方言学会会報』106、昭58・11）は新しい方言区画法「ネットワーク法」の提案である。

林3は他にもいろいろ応用されている。北海道教育大学旭川分校国語学ゼミナール「紋別郡雄武町言語調査報告」（『ことのは』21、昭58・4）では274人のデータを、井上史雄「中学生の八新方言Ⅴ使用パターン」（『国語国文研究』68、昭57・8）では1504人のデータを、荻野綱男「敬語調査への数量化理論第3類の適用」（『言語学演習'83』、昭58・7）では488人のデータを、宮本和美「敬語使用における意識と実際の表現との係わりについて」（『相模国文』9、昭57・3）では540人のデータを、それぞれ林3で分析している。このように林3が普及したことの裏には、コンピュータの普及だけでなく、統計パッケージの普及という事実がある。しかしながら、SPSS統計パッケージで一度に扱える変数の数が100個以下に限定されていることによって分析が規制されるのにはやや不満が残る。また、宮本論文のように、寄与率が低すぎて分析に失敗しているものまで公表されるのは問題であろう。多変量解析法を使えば何でもいいというわけにはいかないのである。

この分野の概説として、荻野綱男「方言研究とコンピュータ」(『方言研究年報・続7』、昭57・12)があった。

三、テキストデータの計量的研究と計量語彙論

この分野でも国研の仕事が目立つ。国研『高校教科書の語彙調査』(秀英出版、昭58・3)は60万語にのぼる大規模な語彙調査の結果である。国研『現代表記のゆれ』(秀英出版、昭58・3)は新聞調査のデータを利用している。斎賀秀夫(編)『話しことばの計量国語学的調査・分析のための基礎的研究』(科研費報告書、自家版、昭58・3)は、『言語生活』誌の「録音機」欄などの話しことばの処理である。中野洋「流行歌の語彙」(明治書院『講座日本語の語彙7 現代の語彙』、昭57・11)は題材として流行歌を取り上げているが、計量語彙論の論文である。

古典テキストもコンピュータ化されつつある。国文学研究資料館(以下国文研と略称)では独自のフォーマットのデータベース化をすすめている。国文研「古典テキストデータ用データベースシステムの開発」(『国文研報告』11、昭58・3)にくわしい国文研方式フォーマットの解説がある。いろいろなレベルの単位切り・属性づけができるようになっている。その他の報告としては、市古貞次(編)『国文学語彙検索システム及び索引誌の作成に関する研究』(科研費報告書、自家版、昭57・3)、田嶋一夫「国文学データベース形成と日本語処理上の課題」(『文部時報』1270、昭58・3)などがある。国文研以外では、東京大学グループの活動がある。風間喜代三他「『ぎやどぺかどる』の読解に於ける電子計算機の利用の試み」(科研費報告書、自家版、昭58・3)は実用的・本格的な古典テキストのコンピュータ処理である。また、計算機利用言語学研究会『言語研究の中の計算機』(自家版、昭57・12)や、荻野綱男・古田啓「人文系研究のための言語データ処理入門」(朝倉書店『朝倉日本語新講座6』、昭58・11)には、これからコンピュータを使おうとする人達へのアドバイスになるようなことがたくさん書かれている。その他、西端幸雄「マイコンによる索引作り」(『椋陰国文学』21、昭58・11)では「拾遺和歌集」をカタカナで入力している。

田中康仁・野村雅昭「サ変動詞の抽出と分析」(『計量国語学』13—4、昭57・3)と田中康仁他「科学技術文献抄録における片仮名列の解析」(『計量国語学』14—1、昭58・6)は科学技術文献の磁気テープから機械的・大量に特定の文字列を切り出し、KWICを作り、計量的な分析を加えるものである。

計量語彙論の分野では水谷静夫『語彙』(朝倉書店『朝倉日本語新講座2』、昭58・4)が光っている。各種の応用もさることながら、「用語データの作り方」を懇切丁寧に説いた点でたいへん参考になる。また、水谷静夫『数理言語学』(培風館、昭57・1)は計量語彙論を中心としつつ広く文法や意味を「数理的観点」から見たものである。山崎誠「文章の話題の展開を計る尺度」(『計量国語学』13—8、昭58・3)は、ある区間の先行区間全体に対する非類似度を用語類似度Dを利用して計るものである。水谷静夫・松盛千佳「同一素材歌謡曲の用語類似度」(『計量国語学』13—4、昭57・3)は、水谷の指標Dで類似度を求め、林4を適用している。

この方面でのコンピュータの利用も盛んになったと感じる。

四、自然言語処理のための日本語研究

工学系の人達による日本語処理の発展に伴って、コンピュータで扱えるような明示的な文法論・意味論が望まれるようになった。この方面では、二つの動きが目立った。

一つは、水谷静夫が編集する朝倉日本語新講座である。すでにふれたもの以外に『文法と意味Ⅰ』『運用Ⅰ』『運用Ⅱ』の三巻が出た。いずれも「手続化志向の態度で臨む」という編者のことばどおり、日本語情報処理を念頭において書かれたもので、人文系の人達の執筆によるものの、工学系の人達にも役立つであろう。

もう一つは情報処理振興事業協会（ＩＰＡと略称）である。『ソフトウェア文書のための日本語処理の研究』シリーズ（昭和57年3月に1号、昭和58年12月に3号から5号）が出ている。1号と5号はコンピュータで扱えるような日本語の辞書を作るための基礎的な研究であり、比較的若い人達による新しい観点からのフレッシュな論考が詰まっていて、着実にデータを集めつつあるようすがうかがえる。3号と4号は「情報処理技術者のための言語学入門」と題したもので、郡司隆男がＩＰＡで行った講義の速記録をもとに書き直したものである。生成文法・文脈自由文法・形式論理学・モンタギュー文法それにＧＰＳＧの解説である。

以上の二つに共通していえることは、主として工学系の人達によってすすめられてきた日本語情報処理の研究が、単なる文字面の操作に留まらず、深い言語理解をめざしており、人文系の研究者に言語の全体像を明らかにするように要求しつつあるということである。おいしいところをツマミ食いしたような「論文」では、彼らの要求に答えたことにならない。

以上の研究と関連して、石井久雄『動詞に対する格の顕現』（科研費報告書、自家版、昭58・3）はパソコンで資料整理をして、個々の動詞に対してよく使われる格は決まっているということを示しており、興味深かった。

五、日本語情報処理と自然言語理解の研究

現在最も活発に活動している分野である。情報処理学会の中の自然言語処理研究会では、この2年間で12回の研究会、合計88件の発表があった。それにしても大変な数である。また、昭和58年6月には「自然言語処理技術シンポジウム」が2日間にわたって開催され、文解析6件、機械翻訳6件、テキスト分析7件の計19件の発表があった。これらをすべて紹介する余裕はないので、筆者が感心した一件だけを取り上げよう。宮崎正弘他「日本文音声出力のための言語処理」（自然言語処理技術シンポジウム）は電電公社横須賀通研の発表だったが、30万語の辞書を用意し、それを用いた高品質の単語分割を行い、接続関係を求めるために３００種以上に品詞カテゴリを分類し、ひらがな列・漢字列に対する高度な複合語処理を行い、同形異語の読み分けのために人名・地名なども辞書に登録して係り受けをくわしく解析し、数詞・助数詞・未知語の読み方のアルゴリズムを指示し、連濁・アクセント・ポーズの処理まで行っている。こうしてコンピュータが文章を「読める」ようになったわけである。人間なら簡単に読めるものでも、その読み方をコンピュータに教えるとなるとことほどさようにむずかしい。ある意味での「読み」の総合的研究が工学系研究者によって行われているわけであ

る。

自然言語理解の研究は言語学・心理学・計算機科学などの学際的研究領域であり、「認知科学」と呼ばれている。この分野での野心的な入門書として田中穂積他『ＬＩＳＰで学ぶ認知心理学３　言語理解』(東京大学出版会、昭58・7)がある。初歩からはじまって流れを追っていく段階的な解説に加えて、具体的なプログラムとそれを説明する豊富な図が載っており、大変わかりやすい。また、制限された辞書と文法ではあるが本格的な言語理解システムＴＱＡＳを提示し、この分野の現在の問題点を述べている。これから取り組もうとする人にとって必読の書である。田中穂積・石崎俊「コンピュータによる談話認知」(『月刊言語』12―12、昭58・12)は自然言語理解における談話認知を扱っている。安西祐一郎「コンピュータに文章は書けるか」(『言語生活』374、昭58・2)はコンピュータの記号処理メカニズム・概念依存文法理論・意味の扱いと機械翻訳などについて述べている。山梨正明「意味と知識構造」(『数理科学』240、昭58・6)は意味的知識を表現するモデルについて生成文法の立場から述べている。参考文献が充実していて有用である。この方面の概説書としては淵一博他『認知科学への招待』(日本放送出版協会、昭58・10)などがある。

機械翻訳の研究はますます盛んになっている。西田豊明他「モンテギュー文法に基づく英文和訳システムの試作」(『情報処理学会論文誌』23―2、昭57・3)はＬＩＳＰ1・7によるものであり、長尾真他「科学技術論文標題の英和機械翻訳システム」(同)は、機械翻訳の対象として比較的やさしい「論文標題」の特徴をうまくつかんでいる。長尾真(編)『機械翻訳システムに関する基礎的研究』(科研費報告書、自家版、昭58・3)は既発表の論文を集めた部分が多いが、それだけで一書をなすことができるほど著者らの活動が活発だということを物語っている。野村浩郷「機械翻訳」(『情報管理』25―9、昭57・12)は概論的だが、具体的な記述で、各機関のシステムをサーベイしている。石綿敏雄「機械翻訳における意味と構文」(『日本語学』2―12、昭58・12)は格文法の問題点を述べ、スクリプトと概念依存文法を紹介し、その必要性・有効性を認めている。

構文解析についての研究も多かったが、その中で特徴的なものとして、吉村賢治他「文節数最小法を用いたべた書き日本語文の形態素解析」(『情報処理学会論文誌』24―1、昭58・1)は従来の最長一致法とちがった新手法を提案している。

辞書をめぐっての話題もあった。長尾真(編)『言語辞書活用のための計算機プログラミングシステムの開発と言語辞書の解析』(科研費報告書、自家版、昭57・3)は『新明解国語』・『コンサイス英和』などの辞書が計算機で扱えるようになったという話である。木村泉「日本語文入力用カタカナ語検出規則とオンライン国語辞典の一分析」(『情報処理学会論文誌』23―2、昭57・3)は『新明解』の見出しを資料にカタカナ語(外来語)とひらがな語を区別する方法を論じている。稲永紘之・吉田将「日本語処理のための機械辞書」(『情報処理』23―2、昭57・2)は九州大学グループが蓄積してきた8万7千余語の辞書を研究用に公開するとのことである。岩波や小学館のように国語辞典を電算写植で作成しているところも辞書を磁気テープの形で販売するようになってきたので、日本語情報処理とか計量的研究などに広く利用されるようになろう。

この分野では、長尾真『言語工学』(昭晃堂、昭58・6)という入

門書が出た。この本は工学系の人の執筆したものではあるが、言語理論や言語の統計的性質からはじまって、言語情報処理の技術的側面についても解説し、文法や意味についてもその処理のしかたについて論じている。

六、数理的研究全般・その他

数学者の立場から言語を見たものもあった。野崎昭弘『数学屋のうた』(白揚社、昭57・4)は、数学者から見た数理言語学というような色彩の既発表の論文集である。著者の幅広い視野が感ぜられる。有川節夫「ことばの数学」(九州大学出版会『九州大学公開講座5　ことばの科学』、昭58・3)は文脈自由文法・オートマトンを平易に説明している。

数理的研究全般に関係するものがいくつかあった。草薙裕『コンピュータ言語学入門』(大修館、昭58・3)、同「私の計算言語論」(『日本語学』2―5、昭58・5)では、コンピュータ言語学を「コンピュータで処理することを前提とした言語の研究」であり、機械翻訳・自然言語理解などのための応用言語学の一つととらえている。特に前者は現在普及中のパソコンに焦点をあわせており、タイムリーなものになった。石綿敏雄「私の数理言語論」(『日本語学』2―5、昭58・5)は言語の理解と表現の数理的モデルについて述べたものである。また、中野洋「コンピュータ言語学」(『言語生活』379、昭58・7)はこの分野全体の総合的解説であり、参考文献が充実している。

計量的研究全般を扱うものとしては、林大他(編)『図説日本語』(角川書店、昭57・2)がある。広範囲の計量的研究の概観ができると同時に、巻末の「ことばの統計学入門」が有用である。

七、コンピュータの発達と今後の数理的研究の方向

パソコンの普及にはすさまじいものがある。以上見てきたような多くの研究の他に、本稿では触れなかったが、まだまだたくさんのパソコン利用があった。こうして各種のデータがコンピュータの中に蓄積されていけば、自然と数理的研究の発展にも結び付いてこよう。たしかにパソコンは便利だし手軽である。しかしソフトウェアの充実していない機種を購入し、本来パソコンに付属していてもいいような基本的なソフトさえも自作するのでは、本当に便利なのかどうかわからない(本人にとってはおもしろいだろうが)。言語研究者として、もっと他にやるべきことがあるだろう。また一部にはパソコンに過大な期待をいだくむきもあると聞く。今後は大型機とパソコンとを使い分ける態度も必要になってくるのではないだろうか。

主として工学系の人達が行ってきた日本語情報処理の研究は、コンピュータの発達にともなって、これからますます盛んになっていくだろう。彼らの研究はしだいに総合的言語研究になりつつあるが、それとともに彼らは人文系研究者による言語研究の一層の発展を求めている。われわれは、従来の狭いカラに閉じこもるのでなく、彼らの期待に答えていく態度でありたいと思うものである。

この展望は、数理的研究とみなされるすべての著書・論文を網羅したものではなく、筆者の興味に合わせておもしろいと思うものを中心に取り上げたので、もれの多い不完全なものになっている。お許しを願いたい。本稿執筆に際し、国研図書館・同情報資料室にお世話になった。ここに記し、感謝する。

――埼玉大学講師――