（「国語学」153集　1988・6・30）

数理的研究

山崎　誠

一、はじめに

日本語の数理的研究は、現在の国語学の中ではやや特殊な位置を占めている。それは、言語現象の分析を数理的に行うものであって、文法研究や音韻研究と違って対象となる言語現象が限定されている訳ではないからである。

数理的研究といっても、用例数の多少など、数を分析に用いている研究は膨大にある。あるテキストでの類義表現の使用回数を調べて云々程度のものを数理的というには、やや幼稚すぎる。筆者は、そのような幼稚な数量表現による分析は方法論的に言えば、数理的以前の常識的段階であろうと考える。

過去の展望号を見ると、この稿に期待されているトピックスは、計量語彙論、数理文法論、統計的文体論、敬語の数量化、機械翻訳などの言語情報処理であろうが、計量語彙論は、語彙で、数理文法論は文法で、……というようにそれぞれの対象である項目で扱い、その他は、「応用日本語学」とでもいう分野にまかせて、もはや「数理的研究」の立項はやめるべきであろう。日本語を対象とした応用言語学は、日本語教育・日本語情報処理を筆頭にますます拡大してゆくと考えられ、認知科学的な研究も増えてくるため現在のような項目立てでは、国語学者の関知できる都合の良い応用分野だけが特別扱いされている点で、中途半端である（応用分野はさっぱり切り捨てるか、積極的に取り上げるか明確にした方が良い）。

また、展望に期待されているのは、個々の文献の列挙では無く、研究動向やその根底にある思想・思潮のようなものを批判することであろうと推察する（文献の網羅は、『国語年鑑』や国語学会・国立国語研究所で編集している『国語学研究文献総索引』に任せるべきであろう）。して見ると、二年という期間で、それほど変化があるだろうかという気もするが、いわゆる「数理的研究」に関しては事情は別のようである。

昭和六一年・六二年は、日本語を計算機で処理するための工学的研究がいよいよ盛んになった期間であった。人工知能の研究の進展とあいまって認知科学が注目された。応用言語学として（時には殆ど純粋工学的な問題として）の日本語研究が隆盛であった一方、言語現象を数理的に分析・解明する研究は、社会調査型の研究で数量化が積極的に行われたことをのぞけば、さほど進展は見られなかったようである。国語学の資料志向の傾向はこの分野にも見られ、データ

ベースや機械辞書に対する関心が強くなった。

二、言語学と言語工学

言語研究者と言語情報処理研究者との交流が盛んになるにつれて、それぞれの分野の根本姿勢を明確にしてその違いを認識する必要が出てきた。例えば、工学者は大部分の説明がつけば例外は無視しても良いと考え、乱暴な理論を立てるが、国語学者はその例外にのみ着目したがるのだそうである（「日本語処理の立場から国語研究に望む」（『計量国語学』15―8、昭62・3）。

言語工学という名称が浸透してきているが、これと言語学的言語研究との間に一線を画す必要がある。例えば、ワープロの機能として「かな漢字変換」というものがあるが、その効率が悪いと、致命的な欠陥とみなされるため、民間を中心に効率の良いかな漢字変換の研究が盛んである。しかし、その研究は特定の技術をまっとうするための研究であって、その必要性はその技術の存在に依存している。つまり、ワープロのような機械が生まれなければ、かな漢字変換の研究の必要性も無い訳である。特定の機械技術に制約されて在る言語工学的研究は、機器環境の進歩とともに新しい研究側面が出てくるといううまみもあるが、技術革新にふりまわされた使い捨て的な研究を生む結果となろう（電子計算機導入当初の国語研究所のいわゆる計量的研究が現在では殆ど記念碑的価値しか持たないことを思い合わすべきである）。言語学的言語研究は、「日本語での同音異義語が多いのはなぜか」とか「一つの語が複数の表記を持つ場合の使い分けの条件」など言語現象の本質に対する興味に根ざしている。ワープロのかな漢字変換機能の評価を日本語研究の立場から分析し、「かな漢字変換とは、語の認定のことであるから、前後の文脈が分からなければ決定できないだろう」などの提言をすることは出来よう。しかし、言語学的言語研究の観点に立てば、そのような応用面は無視できるし、積極的に区別をすべきであろう。（勿論、一人の人間がその両方を兼ねられないことはないから、言語学者であり、言語工学者であっても良い）。言語工学を基盤として言語産業が成立すると考えられている（長尾真「機械翻訳システムの基本的な考え方」（『日本語の特性と機械翻訳』、出版科学総合研究所、昭62・7））。おそらく、人文系日本語研究者がそれに係わる傾向は今後増大するであろう。

三、計量的研究

最近の計量的研究の傾向は、被験者を使った実験・調査型の研究と、実態把握・記述型の研究との分化が鮮明になってきたことである。両者の違いは、実験・非実験にあるのでは無く、扱っている言語現象を、言語意識の問題としてとらえているかどうかの違いによる。

水谷静夫氏の還暦記念論集として『計量国語学と日本語処理―理論と応用―』（秋山書店、昭62・3）が刊行された。数理言語学・計量国語学で統一された論集というのは、日本ではこれが初めてであろう。（以下、『計日』と略す）。

言語研究と計量的方法の関係を論じたものに、荻野綱男「計量言語学の立場と方法―「理論言語学」への疑問―」（『理論言語学』（特定研究「言語情報処理の高度化のための基礎的研究」昭和62年度第4回研究会要旨集））がある。これは、荻野自身が近年強く打ち出している姿勢で、言語学の範囲をいわゆる応用言語学などをも含んで広くとら

え、言語を人間行動の一つとして考え、現実のデータからの帰納・モデル化・実験を通して、人間の言語のしくみを解明しようとする立場である。

文法…アンドレイ・ベケシュ『テクストとシンタクス―日本語におけるコヒージョンの実験的研究』（くろしお出版、昭六十二年一月）は、最短文（テクストを構成する要素）間のパラフレーズ調査を通して文章のコヒージョンを計量的手法で解明しようとした。宮島達夫「格支配の量的側面」（『論集日本語研究(1)現代編』明治書院、昭61・11）、同「格の共存と反発」（『計日』）では、空間移動動詞の結合価の実態の量的な違いを調査し、複数の格をとる場合の格どうしの共存、非共存の傾向を統計的にとらえている。

表記…佐竹秀雄「漢字・平仮名・片仮名の比率の関係」（『計日』）は、最近の文章における片仮名の役割を漢字・平仮名と比較して明らかにしたものである。徳田克巳「漢字の読みの難易性評定と漢字属性の関係」（『読書科学』121、昭62・10）は、漢字の読みの難しさに影響を与える属性は、画数や具体性、熟語数などであることを実験的に示している。

語彙…最近の計量語彙論的研究の関心のひとつは、国立国語研究所『雑誌用語の変遷』（秀英出版、昭62・3）に象徴されるように歴史的な変遷をとらえることにあるようだ。同書は、語種・品詞の変遷、意味分野ごとの語誌が記述されている。また、土屋信一「江戸語会話文の漢語使用率」（『計日』）、同「浮世風呂・浮世床の会話文の漢語使用率」（『国語語彙史の研究』八、和泉書院、昭和62・11）は、江戸語を中心に漢語使用率とその変遷を調査したものである。真田信治「理解語彙量の累増過程―ことばの習得をめぐる事例研究―」（『日本語・日本文化研究論集』（大阪大学）、昭61・1）は、同一個人の理解語彙量を六才時と十才時とで比較し、その間の増加の様相を品詞、意味範疇ごとに記述したものである。テーマは違うが、田中章夫「近代の表記における漢字依存度の変遷」（『計日』）も変遷を扱ったものである。その他では、田中章夫「語彙研究における順位の扱い」（『国語語彙史の研究』7、和泉書店、昭61・12）は、主に古典作品の統計資料に基づいて頻度、比率、共通度についての順位の相関のとらえかたを示唆したものである。中野洋「話しことばの語種の調査」（『計日』）は、『言語生活』誌の「録音器」欄データを基に語種と比率と話し手の属性（性、年齢、職業）との関連を調べたものである。

意味…荻野綱男「語の意味の対立と文の意味の対立―新しい「計量意味論」の提唱―」（『日本語学』6―7、昭62・7）は、語や文についての反対（語）意識の数量化を試みた。山崎誠「類義性の数量化か可能か？」（『日本語学』5―11、昭61・11）は、語の類義性の数量化の限界を示したものである。SD法によって言語的直観や言語感覚を数量化する試みとしては、馬場俊臣「文中の語の相互関連度に対する言語的直観の研究」（『計量国語学』15―5、昭61・6）、竹内晴彦ほか「温冷感・乾湿乾に関する言葉の意味の分析」（『計量国語学』15―6、昭和61・9）がある。前者は、文中の語と語・文全体の凝集度（cohesion）を測定しようとする試みである。

方言研究…方言区画を計量的に決めるための試みとして、柴田武・熊谷康雄「ネットワーク法における地点間の言語的類似の新しいとらえかたと処理のしかた」（『国語学』150、昭62・9）、井上史雄「文法現象による計量的方言区画」（『言語研究』89、昭61・3）があげられる。前者は、地点間の言語的類似をそれぞれの地点が持つ類似的関

係に着目して数量化したもので、著者らの従来の方法を改良したものである。

敬語…社会的諸属性と言語使用との関係を探る調査的研究においては、林の数量化理論の利用など、高度な数量的方法による分析がごく普通のようになりつつある。伊藤隆「若い世代の非標準的表現と使用者の社会的特徴」(『国語学』147、昭61・12)は、いわゆる「新方言」の使用の要因を林の数量化Ⅲ類を用いて分析し、性別・テレビ視聴・個人のネットワークの三つに要因を求めている。吉岡泰夫「高校生の敬語知識とその形成要因」(『計量国語学』15―6、昭61・9)、同「敬語行動における知識・態度・意識・使用」(『熊本短大論集』38―1、昭62・8)は、敬語意識や敬語使用が敬語知識にどう関与しているかを林の数量化Ⅰ類を用いて分析したものである。

文体論…石純姫「多変量解析による文体分析―孝標女をめぐって」(『中央大学大学院論究』19―1、文学研究科篇、昭62・3)は『夜の寝覚』『浜松中納言物語』を基本的な文体統計量に基づいてその作者を菅原孝標女であろうと推定したものである。

四、データ・資料

国語研究所は、コンスタントに語彙調査の結果を出し続けた。『中学校教科書の語彙調査』(昭61・3)同 II(昭62・3)、『雑誌用語の変遷』(昭62・3)である。

教科書調査は、理科系・社会科系教科書の長・短2種類の単位での語彙表である。延べ語数はそれぞれ、約二十万、約二十五万である。『雑誌用語の変遷』は「中央公論」の経年語彙調査で、一九〇六年からの十年間隔で、一万語ずつのサンプリング調査である。各種語彙表の他に語彙を中心に雑誌用語の変遷を記述している。また、マイクロフィッシュの形で、国立国語研究所言語処理データ集『高校教科書―文脈付き用語索引―』も刊行された。

次の五に述べる特定研究の成果として、以下のような資料集が刊行されている。

田中康仁(作成)『語と語の関係解析用資料―〝を〟を中心とした』(3分冊、昭62・3)、同『語と語の関係解析用資料―朝日新聞記事データ分析―〝を〟を中心とした―』(3分冊、昭62・11)。柴田武『『分類語彙表』と『新明解国語辞典』のマッチング・リスト』(昭61・12)。野村雅昭・石井正彦『学術用語語基連接表』(昭63・3)。前二者は、「名詞+を+動詞」の連接表である。それぞれ日本科学技術センターの抄録文と朝日新聞記事データ(84日分)を加工したもので、用例数はそれぞれ、約十八万、約十万である。『学術用語語基連接表』は、文部省編『学術用語集』二十三編の全収録語を構成する語基(語構成の単位)の一覧表であり、結合した語形及び使用学術分野が示されている。収録語数は約十五万である。

後述のように、言語情報処理の研究者は基本的・網羅的な言語データの提供を求めているようだが、データ供給側の思惑とうまくかみあっていないように思える。双方の歩みよりが必要であろう。

五、日本語情報処理の研究
(機械翻訳・データベース・シソーラスなど)

文部省の科学研究費特定研究「言語情報処理の高度化のための基礎的研究」が昭和六一年度からスタートした。理論言語学・対照言

語学・文章理解・言語データ解析・情報ドクメンテーション・認知心理学などの研究者が多数参加した大きなプロジェクトである。その成果報告として「言語情報処理の高度化のための基礎的研究総括班研究成果報告書」1986（昭62・3）、同1987（昭63・3）が出ている（以下、『高度化』と略す）。また、言語情報処理の研究が雑誌に特集のテーマとして好んで取り上げられた。『日本語学』昭六一年六月号「言語情報処理の言語学」、『言語』昭六一年七月号「高度情報化社会のことば」、『情報処理』昭六一年八月号『計算言語学』、『言語生活』昭六二年一月号「機械がとらえることば」、『言語』昭六三年一月号「機械翻訳の現状と未来」などである。

また、次のような言語情報処理関係のシンポジウムも開かれた。「日本語の特性と機械翻訳」（第1回「大学と科学」公開シンポジウム。同報告書、出版科学総合研究所、昭62・7）、「日本語処理の現場から国語研究に望む」（計量国語学会、同特集号、15―8、昭62・3）。

機械翻訳は、その限界がはっきり認識されて来たといってよいだろう。日常言語のような（機械側に言わせれば）省略の多い曖昧な表現や文学作品は扱わず、科学技術文献・技術文書・事務書類のような明解な文章を対象とするという方向が強く認識されてきた（辻井潤一「機械翻訳―現状と未来―」（『言語生活』422、昭62・1）、長尾真『機械翻訳はどこまで可能か』（岩波書店、昭61・2））。最近では機械翻訳と言うよりも機械翻訳支援システムなどというように支援という言葉を補う傾向にあるそうである。

言語情報処理の研究者と言語学の研究者との交流が多くなったためか、言語学と言語情報処理との関係や相違点、お互いへの要請が論じられた。機械翻訳に対して言語学がどのような貢献ができるかという点について、郡司隆夫「言語学と機械翻訳」（『言語』17―1、昭63・1）では、翻訳技術は言わば大工の職人芸的なノウハウであるから、言語学の理論構成の関知するところではない。機械翻訳が成功するかどうかは、言語学の理論の完成ではなく、翻訳技術が計算機上の技術に移植できるかどうかだと主張している。一方、「日本語処理の現場から国語研究に望む」（『計量国語学』前出）では、「言語学者はきれいな文法規則を出すけれど、実際に処理すると例外が多くて困っている」「国語研究者が、基本的で網羅的なデータをもっと提供してほしい」などの意見が出されている。後者のような要望は、以前なら国語学者の方から言語情報処理研究者に出されてしかるべきものであろう。また、長尾真「言語情報処理から言語学に期待するもの」（『日本語学』6―5、昭62・5）では、言語現象全体のメカニズムの説明、総合的言語モデル、複数の言語の対照研究の具体的成果などを言語学に対する要請としてあげている。そのためには、従来の言語学の枠組みを越えて、認知科学的立場が重要だと指摘している。

機械翻訳においては、前処理ということが言われる。自然言語は曖昧だから機械翻訳にかける前に省略を補ったり、係り受けを明確にしたりして構文を整えてやらなければならないのである。機械翻訳はコストで評価される面が強いので、前処理・後処理で人間が手を加えたとしても全部人間が行うよりも能率が良ければ良いのである。このような発想が進むと、機械処理に都合の良いように、日本語の規格化（controlled Japanese）という制限言語の考えも提唱されて来る（吉田将「日本語の規格化と制限日本語の設計」（『日本語の特性と機械翻訳』前出）。これは新たな国語問題を引き起こす可能性をはらん

でいると言えよう。

「データベース」という言葉が国語学者の間でも使われるようになってきた。その理解の仕方はテキスト・データベースが主かもしれないが、利用価値の面から、必要性だけはよく論じられている。データベースの作成自体は全く研究でも何でもなく、単なる作業である。これを研究であると誤解している国語学者はあまりいないようであるが、データベースの構築・利用の仕方やその中身の構造を考えたりして総合的にデータベースを研究することが言語の数理的研究の一分野であると思われている節もある。しかし、それも、本来は、情報処理研究・ドクメンテーション研究に属するものであろう。

機械処理の必要から、機械辞書・計算機上でのシソーラス作成の試みも活発に行われた。機械辞書では、中野洋ほか「日英語彙データの収集・比較と機械辞書の作成」(『高度化』1986、1987)、鶴丸弘昭「日常辞書の機械化とその応用」(『日本語の特性と機械翻訳』前出)(『新明解国語辞典』の機械化)、荒木卓也・栗原茂信「多面的検索機能を備えた機械式辞書の試作」(『計量国語学』15―6、昭61・9)(読み、画数、部首、四角号碼、SJIS コードを組み合わせて検索できるようにした漢字表記語辞書のシステム)などがある。機械辞書の諸問題を論じたものに、林大ほか「機械辞書の時代」(シンポジウム記録、『言語情報処理の高度化研究報告3 ディスコース・機械辞書』昭62・6)、村田賢一「計算機用動詞辞書に於ける意味記述の試みについて」(『計日』)、田嶋一夫「漢字辞書の構成」(朝倉日本語新講座1『文字・表記と語構成』朝倉書店、昭62・12)(計算機で漢字辞書を実現するために必要な種々の統計的属性、漢字データ処理上の問題点(コード、異体字、検索方法、外字処理)をまとめたもの)、佐竹秀雄「表記辞書と仮名／漢字変換」(前出『文字・表記と語構成』)。

シソーラス関係では、荻野孝野「日本語の意味分類試案」(計量国語学会第31回大会発表要旨、昭62・9)、同「日本語の意味分類体系」(『計量国語学』16―3、昭62・12)、田中穂積「上位／下位関係シソーラスISAMAP1の作成、I、II」(『自然言語処理』64―4、昭62・11)、荻野綱男ほか「現代日本語の名詞シソーラスの作成」(『高度化』1986、1987)、荻野綱男「シソーラス作成の問題点」(『日本語学』6―5、昭62・5)など。

研究支援ツールに関しては、ワープロの評価やパソコンソフトの作成が多かった。荻野綱男「カナ漢字変換システムの能力の調べかた VJE―βを例にして」(『計量国語学』16―2、昭62・9)、樺島忠夫「日本語ワードプロセッサへの提言」(『前出『文字・表記と語構成』)。西端幸雄「マイクロ・コンピューターによる自動品詞認定の試み」(『樟蔭国文学』23、昭61・1)(形態的特徴を手掛かりに語の品詞認定を行うシステム)、同「文字処理のためのサブ・ルーチン」(『樟蔭国文学』24、昭62・3)、松尾雅嗣・鈴木重樹「パソコン用簡易用例検索プログラム―PC―KWIC―」(『計量国語学』15―4、昭61・3)、伊藤雅光「パソコンによる漢字仮名混じり文用KWIC索引作成システム―『こもれび』―」(『計量国語学』16―3、昭62・12)。この他にも、パソコン用ソフトは個人的なものが多数作成されていると思われるが、汎用性の高いシステムの開発・紹介が論文として通用する段階からはもう脱してもよいのではないか(これは、研究効率の良し悪しとは別の問題である)。「計量国語学」誌の内容がその意味で最近安易に流れているような印象を受ける。

六、最後に

筆者の力不足のため、見落とした文献あるいは、内容の誤解等があるかもしれない。また、独自の研究観が出すぎていて客観性に欠けているかもしれない。これらの点での非礼は全て筆者の責任である。

――国立国語研究所員――