コンピュータ言語学

国立国語研究所の歩み・1

コンピュータ言語学

国立国語研究所
昭和48年6月

国立国語研究所の歩み・1

コンピュータ言語学

一　コンピュータ言語学の起りとその大勢

コンピュータが出現してから、それまで人間だけが使っていた言語をコンピュータも使うようになった。人間だけが言語を使っていたころ、言語学は、人間の言語を分析することがすべてだった。学問や知識や教養の有無とは関係なしに、人が無意識に使っている言語を分析してみると、そこに驚くべく精巧な組織がある。この組織は見れば見るほど興味のあるもので、音声、文字、語彙、文法などいろいろな面から観察して、記述しても記述しても疑問の尽きることがない。人間の心や頭脳が何とも不思議なものであるのと同様に、人間の言語も不可思議極まるものである。諸方面の科学が急激に発達しだした十九世紀のヨーロッパを中心に、言語学も科学として成立し、二十世紀の世界各国において、言語学は精密科学としての体制を整えるに至った。

二十世紀の後半に電子計算機が出現したことは、人類の歴史上特筆すべき大事件であった。エレクトロニック・コンピュータを文字通りに訳した電子計算機の名称が語るように、コンピュータは数の計算をするために作られた機械である。しかし、この機械は、算盤のような単純な計算器とはちがって、符号化された文字によって情報を記憶し、作業手順を指示したプログラムを記憶装置の中に内蔵することにより、どんな複雑な工程をも間違いなくたどって、それからそれと処理を進め、所期の形で答えを出す。プログラムの中には必ず判断の論理が組み込まれているから、計算の過程において、コンピュータは、二つの情報を比較して異同を判断したり、多数の情報の排列順序を変えたり、雑多な情報の中から必要な情報をさがして取り出したり、ある情報を特定の処理によって変形したりしている。情報にこのような処理を施すことは、その情報が数値である場合にだけ可能なのではない。読み書きのための文字もコンピュータの二進符号になりさえすれば、同様の処理を施すことができる。アルファベット二十六文字は極めて容易に符号化できるから、欧米の言語は早速、コンピュータによっ

て扱われ始めた。日本のいろは四十八文字もコンピュータにとって充分扱い易いものである。人間だけが使っていた言語を、コンピュータも使い始めたのである。

ヨーロッパの諸言語は、十九世紀の比較言語学によって明らかになってきた通り、互いに親類関係にあり、語彙や文法に共通点や特定の関係が見出されている。それならば、Aという国の言語で作られた文は、その単語をBという国の言語の単語に置きかえ、語順を一定の手続きで入れ替えれば、B国語の文に変わるだろう。その変換手順は、かなり機械的なものではないか、ならばその手順をコンピュータに教えることができるだろう。そうすればコンピュータに翻訳の作業をさせることができる、欧米の学者がそう考えたのは全く自然な成り行きであった。コンピュータの第一号といわれるMARKIがハーバード大学に現われたのは一九四四年、そのころからすでに機械翻訳の可能性が論じられ始め、何人かの試みがあったのち、一九五四年には、アメリカ、ジョージタウン大学のドスタートとガーヴィンが、IBM七〇一を用いてロシア語から英語への翻訳に、ともかく成功している。日本の通産省電気試験所（現在「電子技術総合研究所」）で一九五八（昭和33）年に試作されたコンピュータは、初めから翻訳を目的として作られ、その名も「翻訳機ヤマト」とつけられた。このことからも、コンピュータと翻訳とのかかわりの深さが感じられる。機械翻訳の可能性が一度確かめられてみれば、それは、別に、両言語が親類関係にあることとは関係がない。A国語文の構文がコンピュータに理解でき、その内容をB国語の文法に従ってコンピュータが構文化できればよいのである。要は両国語の文法が正確に機械的に把握されることである。

ところが、文法を、機械がデータを処理する手順として、迷う余地のない論理回路に組もうとすると、実にむずかしいことがわかってきた。人間のための言語学は立派に近代科学として成立していたかに見えたが、いざこれを機械に適用しようとすると、ほとんど無力であった。人間は言語を分析する時に、まず表現内容を直観的に理解してから単語と単語の関係を考え、主語だの述語だのに分けて行く。コンピュータは、文字や文字連続を認知することは正確で速いけれども、人間のような意識や直観をもたないから、手がかりなしに内容を理解することができない。コンピュータに理解させるためには、必らず形式上の手がかりを与えなければならない。言語分析の中から直観をすべて排し、ことごとく形式によって判断可能な識別標識づきの処理手順として、文法を組み立てる努力が始った。そのような文法を機械文法とよぶ。機械文法の探求がおそらくコンピュータのための言語学の始まりであろう。

コンピュータによる言語データの処理で翻訳についてすぐに着眼されたのは、情報検索と自動抄録である。情報の海の中から知りたい情報を早く取り出すことが現代生活にとってどんなに必要かは言うまでもない。百科事典や各種参考書類が目次や索引でどんなに整理されていても、知りたいことの答えを見つけ出すのはなかなかひまのかかるものである。まして答えを含んでいるかいないかわからない文献にあてもなく目を通すのはやりきれない仕事である。情報源はみなコンピュータの記憶装置の中に入れておき、キーワードを指定して、必要な事項をさっと取り出すようにしたい。これがコンピュータによる情報検索である。また各種の文献を記憶装置に貯える時に、文献全体の代りに文献中のキーワードやキーセンテンスをピックアップした抄

録を予め作っておけば、情報検索の手間がよほど省ける。その抄録作りを人間がやらないでコンピュータにやらせたら、さらに能率的である。これがコンピュータによる自動抄録である。

情報検索や自動抄録は主に単語を相手にする業務だから、文法を生命とする機械翻訳よりは楽かと思われるが、なかなかそうでもない。インフレのことを知りたいときに、「インフレ」という単語だけを目当てにしたのでは必要な情報をのがすことが多い。必要な情報の所在が「物価」「値上り」「デフレ」「生活難」など、どんな単語のある処にひそんでいるかわからない。単語は形の存在であるとともに、それ以上に意味の存在である。情報検索や自動抄録を適切に行わせるためには、単語で表わされる意味の体系が明らかにならなければならない。単語同士の意味の関係について、抽象・具体の関係、含む・含まれるの関係、同意語・類義語・反対語の関係、事実との対応から来る、全体と部分の関係、共存・類縁の関係等さまざまな関係が明確にたどれるように記述されなければならない。しかも、よく使われる単語はとかく一語でいろいろな意味をもっていることが多いし、どの単語でも、実際の文脈の中では特殊な意味をもつことがある。各語がどういう文脈の中でどういう意味に用いられる傾向があるかということも組織的に研究する必要がある。情報検索や自動抄録に備えて、機械のために、語の意味の研究が必要になってきた。これらは機械文法に対して、いわば機械意味論ともいうべきものである。機械翻訳も、決して文法だけではできない。ことばのかかり受け関係には、あいまいなものがたくさんあり、その解決に単語の意味の情報が必要になることが多いからである。

以上のように、機械翻訳、情報検索、自動抄録の試みを進める過程で、機械文法や機械意味論の研究が進み始めた。こういう研究は、まず初めは、コンピュータの専門家である工業系の学者の手で行われたが、それらの学者はまず、言語学の研究成果にその拠り所を求めた。二十世紀前半の言語学の中で、最も目ざましい業績をあげたのは、構造主義の言語学である。スイスの言語学者ソシュールの説に端を発するとされている構造主義言語学は、アメリカのブルームフィールドにおいて目ざましい発展をとげ、ハリスにおいて精密な記述法が整ってきた。アメリカの構造主義言語学は、自分たちの言語とは全く構造の違うアメリカ・インディアンの数多くの言語を記述するという課題を負ったために、特に内省や直観を排し、形式だけを尊重して、機械的に分析を進めるという学風を作り出した。その点で、構造主義の言語学には、コンピュータのための言語学に道を開く可能性があった。ハリスの門にチョムスキーが出て生成文法の考え方を発表し、さらに変形文法に発展してくると、この生成変形文法は、機械文法にとって最も有力な理論であることが認められるようになった。コンピュータが実用段階に入ってきたのとチョムスキーの諸論文が大きな影響力をもちだしたのとがちょうど同じ時期の一九六〇年代であることも手伝って、コンピュータによる言語処理の理論は、多かれ少なかれチョムスキーの影響を受けることが必至の状態になった。

生成変形文法は論理学と数学に基礎を置いて、演繹的な考え方をする。単純な基本原理から次々と派生規則を仮設し、現実のあらゆる文が作り出せる規則の集成として文法を組み上げようとする。現実の文に当ってからその説明法を考えるのではなく、まずモデルを作って表現機構を設定し、それによって現実

の文が説明できるかどうかを試していくという考え方をするこの理論は、はじめ伝統的な言語学者には受け入れられなかったが、コンピュータに言語活動をさせてみるのには、この方法が最も適していたので、機械文法としては早くから受け入れられる結果となった。

コンピュータの性能が急テンポで向上していくとともに、コンピュータによる言語処理の試みも範囲が広がってきた。言語音声の聞き取りと合成、文字の認識と発生、漢字入出力の高速化、自然言語による質問応答組織の作成、新聞・雑誌・辞典等の自動編集、各種事務文書の自動作成等が実用化を目ざして進みつつある。コンピュータは大量処理を得意とするから、処理手段が形式的に定められる自動編集のようなものは、単純に人間の労力を省く点で大いに成果をあげているが、自然のままの言語を処理することでは依然として大きな困難をかかえている。理解したり表現したりする人間の言語活動に類似した言語の処理をコンピュータにさせるためには、人間の言語活動をメカニズムとしてとらえるための研究を徹底的に進めなければならない。音声、文字、語彙、文法、文章、文体等、言語のあらゆる面をコンピュータ的視点から研究しなおす必要がある。そういう研究を総称してコンピュータ言語学という。

コンピュータ言語学は、computational lingnistics（CLと略す）の訳語である。「計数言語学」「計算言語学」と訳されることもあるが、「計量言語学」と訳される quantitative linguistics や「数理言語学」と訳される mathematcal linguistics とまぎれやすいので「コンピュータ言語学」の方が明確でよい。「計量言語学」も「数理言語学」もコンピュータの出現以前からあった学問領域である。三者の区別を明瞭に立てることはむずかしいし、区別しても、共通領域が多くなるので、無理に区別する必要もないかもしれないが、明らかなのは、問題をとらえる角度がちがうことである。計量言語学は言語を統計的にとらえ、数理言語学は言語を数理一般の角度からとらえる。ともに、コンピュータとは直接関係がない。コンピュータ言語学は、常にコンピュータ処理との関係において問題をとらえるところに特徴がある。

computational linguistics という名称を始めて用いたのがだれか、明らかではないが、この名が一般化するのに最も与って力があったのはアメリカの言語学者ヘイズであろう。ヘイズは一九六〇年代の大半をランド・コーポレーションの言語学部門主任として過ごし、言語処理に大きな功績を残したが、その間、彼の論文の中にしばしばこの名称が用いられた。今は「国際コンピュータ言語学会議」（International Conference on Computational Linguistics）や、日米科学協力の一環としての「コンピュータ言語学サーベイ・セミナー」（Survey Seminar on Computational Linguistics）などが定期的に開かれるのを始めとして、この名を冠した研究集会や論文著書も多くなった。日本のコンピュータの専門学会である情報処理学会にも、CL研究委員会という分科会があって研究を続けている。

以上述べたように、コンピュータ言語学はコンピュータの使用に関連して行なわれる言語研究を内容とするが、本来、コンピュータという機械の出現によって否応なく起った実用性の学問であるから、その内容は雑多であり、学問として純粋な体系をもっているわけではない。将来、体系ができるかもしれないが、現在は実績によっていろいろな研究をこの名で一括して呼ぶよりほかはない。しかし、その中にもおのずから次の三つの領域が

区別できるように思われる。

1、コンピュータで言語を自動的に処理する方法を求めて言語を研究すること

2、コンピュータを分析の道具に用いて言語を研究すること

3　コンピュータに命令したり、コンピュータと対話するための言語を研究すること

二番目のコンピュータによる言語研究は、例えば、コンピュータでシェークスピアの作品の語彙調査をしてシェークスピアの用語法を研究するというようなもので、この場合、コンピュータは資料作成の道具に使われたにすぎず、研究の目的はシェークスピアの用語法を知ることにあるから、厳密な意味では、この種のものをコンピュータ言語学に含めるべきでないかもしれない。しかし、実際には、コンピュータを用いて言語を調べ、その結果を言語自動処理法の開発に役立てるというコースをたどることが多いので、この項目もかかげておく方が実際的である。三番目のものは、前記の二つとははっきり区別される。コンピュータの命令語を自然言語に近づけようとして作られるものをコンパイラといい、これまでに、COBOL, FORTRAN, PL/1, などのコンパイラが作られた。コンパイラの研究は、本来純粋にコンピュータ内部の研究であり、言語学とは関係のないものであるが、命令語の体系をできるだけ日常の言語に近づけようとするところに言語研究とのかかわりが生ずる。ことに最近は命令語の範囲が拡大されて、専門オペレータでない一般の人が日常の言語で質問したり仕事を命じたりすると、コンピュータがそれを受けて必要な処理をし、出た答えを普通の言語で提供するという質問応答システムの作成までが命令語研究の範囲に入ってきた。こうなると、命令語の研究もコンピュータ言語学に含めるのが適当となる。

コンピュータ言語学とはこのようなものである。以下、コンピュータ言語学の全般ではなく、国立国語研究所が行なってきた研究の範囲内においてその内容を紹介する。

二　国立国語研究所におけるコンピュータ言語学の研究

国立国語研究所では、昭和四十年度にコンピュータの導入がきまり、ＨＩＴＡＣ三〇一〇、二〇ＫＣの機械を設置した。昭和四十一年以来、この機械によって調査研究を続けてきたが、昭和四十八年度から、新しい機械ＨＩＴＡＣ八二五〇、九六ＫＢの機械に更新することがきまった。その間に行なってきた研究の内容を以下に紹介する。前節で述べたコンピュータ言語学の三領域のうち、三番目のコンパイラの研究は行なっていないので、前記二項目に属する研究を紹介する。そのうち、二番目にあげたコンピュータを道具に用いた言語調査の実際をまず記し、次に自動処理法の研究を、直接自動処理にかかわるものと、自動処理を目ざした基礎研究とに分けて記すことにする。

Ａ　コンピュータによる言語処理の実際

言語調査を行なうためのコンピュータは、当然数値計算用ではなくて、事務計算用であるべきこと、言語データは調査単位の長さが不定だから、固定長よりも可変長のデータが扱いやすいこと、データ移動の目標記号を、語頭から語末へ、つまり、左から右への方向でさがす命令が使いやすいこと、などの理由でＨＩＴＡＣを選んだ。この機械を用いて行なってきた業務の主たるものは、新聞の用語用字調査である。

【一】新聞の用語調査

昭和四十一年一年間の朝日・毎日・読売、

三新聞朝夕刊全紙面に含まれる一億八千万語を母集団とし、それから六十分の一の抽出比で抜き出された三百万語を標本として調査した。この「語」は、比較的短かく切られた単語であるが、日本語の単語、殊に漢語は「高速道路建設反対運動対策委員会」などと、いくつもつながって複合語を作る性質があり、文の中ではこういう複合語が一単語の働きをする。文を文節で切って助詞や助動詞を切り離すと、こういう複合語の単語が求められる。この種のものを長単位の単語、または単に長単位と呼んだ。これに対し、短かく切って得られた単語を短単位の単語、または単に短単位と呼んだ。長単位でかぞえれば、母集団は一億二千万語、標本は二百万語となる。今回の用語調査では、文をまず長単位に切ってカウントし、長単位の語彙表を作ってから長単位を短単位に分割し、短単位を集めなおして短単位語彙表を作った。入出力には漢字テレタイプを用いた。サンプルは、エリヤ・サンプリングの手法で紙面のどの部分からも六十分の一ずつ取られた。記事の性格などによってデータを区分し、各語の所属区分がわかるようにした。データ区分は、次の四種の角度から行なった。

①話題による区分　1政治　2外交　3経済　4労働　5社会　6国際　7文化　8地方　9スポーツ　10婦人・家庭　11芸能・娯楽　12広告その他（文種区分の十四から十七までの合併）

②文種による区分　1ニュース　2ニュース解説　3社説・コラム　4時事的特集記事　5特別読物　6評論・論文　7実用知識読物　8探訪ルポ　9長期ニュース展望　10記録・通知　11紹介記事　12読者作文　13相互通信　14小説　15一般広告　16案内広告　17漫画

③署名形式による区分　1無署名記事　2通信社記事　3冒頭署名記事　4外部記者末尾署名記事　5社内記者末尾署名記事　6略称末尾署名記事　7外電冒頭記名記事　8社内無署名記事　9外部記者無署名記事　10広告その他

④位置による区分　1見出し　2標題・欄名　3リード　4本文　5情報源・署名　6表　7写真・図表等の説明　8広告その他

このように、四種類もの角度からデータ区分をしたのは、次のようなことを考えたためである。

(ア)、新聞は、話題からいっても文章形式からいっても、雑多な要素を含んでいるから、それらを分析区分し、それぞれにおける用語の特徴を見出したい。文法形式や文体上の特徴も、これで、ある程度わかるだろう。

(イ)、基本語彙を見出すためには、全体での使用度数だけでなく、部分部分での使用度数を知り、どこでも広く用いられる語とどこかで特別に用いられる語とを分ける必要がある。

(ウ)、情報検索や機械翻訳の効率を高めるためには、文章一般において、近い文脈の中での語の共存関係を調べる必要がある。文章の種類によって共存関係の型に違いがあるだろう。

(エ)、コンピュータは、分類標識を与えておけば、大量のデータを簡単に分類することができる。一種類だけの分類で満足しては、もったいない。（⇩文献7の林論文、木村論文、文献8の林論文）

昭和四十七年に全データの長単位調査を終了した。その結果、概算延べ二百万の実数は百九十六万余であった。これに含まれた異なる語の数は約十九万である。この中には、単語としては無意味な経済株式欄の数字や記号などが多量に含まれている。そういうものを除いた十六万余の長単位語彙表を四十八年春に刊行した。（⇩文献4）それまでに、三分

の一を処理した段階で、長単位および短単位の語彙表を三冊刊行している。(⇩文献1、2、3)

これら語彙表における単語の排列には、五十音順と使用度数順とがあるが、五十音順に二種類がある。真の五十音順と疑似的な五十音順とである。疑似的五十音順というのは、すべての漢字にその字を代表する一種類の音を与え、語頭に立つ漢字の代表音によって排列したものである。最初に入力した新聞の文章は、よみがなのついていない原表記のままのものであるから、コンピュータが長単位のデータを扱う段階では漢字のよみ方がわからず、従って五十音順に並べることができないので、便宜的な方法として、この疑似的な五十音順を用いた。(⇩文献6の田中論文)長単位を短単位に切る時によみがなをつけたから、短単位の表は正しい五十音順になっている。

【二】新聞の用字調査

用語調整と同じデータを用いて文字・表記の調査を行なっている。これは次の三種の調査に分れる。

①漢字の使用度数調査

②漢字の用例調査

③語表記のバラエティの調査

①の漢字使用度数調査は最も簡単に施しうるもので、三分の一データに含まれていた二千九百八十種の漢字についての中間報告を刊行してある。(⇩文献5)この漢字表には、使用度数5以上の漢字二千百四十三字については、総度数のほかに、用語調査の話題区分に大体準じた十二区分によって、記事区分別使用度数が示してある。

②の漢字用例調査は、各漢字について、一般用法と固有名詞用法とを分け、それぞれの中を音訓別に分けた上でまず短単位用例を度数つきで列挙し、各短単位のあとにそれを含む長単位用例を度数つきで示すものである。この調査では、素集計までをコンピュータで行ない、あとは人力によって原文にもどり、正確なよみがなをつけた。全資料の中から広告を除いて得られた資料について現在作業中である。これが完成すると、国語国字問題を考えるための最も有力な資料になるばかりでなく、コンピュータで日本語のデータを処理する上での重要な資料となる。

③の語表記バラエティ調査は、例えば「アゲル」という語について「上げる」「揚げる」「挙げる」「あげる」などの書き分けの数量的実態を明らかにするもので、これも②の場合と同様の意義をもつ。同じく現在作業が進行中である。

【三】漢字入出力の方法

HITAC三〇一〇は6ビット・コードを用いている。6ビットで区別しうる文字の種類は六十四であるから、もちろん漢字を扱うことはできない。現在のコンピュータは大部分8ビット・コードを用いるようになったが、8ビットで区別しうる数は二百五十六だから、これでも漢字は扱えない。用語調査だけなら、かな文字で入出力してもよいが、漢字を調べることが大きな目的である場合には、どうしても漢字は漢字のままで扱わなければならない。そこで前記の用語用字調査の入出力には、漢字テレタイプを用いた。漢字テレタイプは、新聞社が活字の自動鋳造と記事の遠隔地送信とに使っているもので、6ビット2列の12ビット(これで区別しうる数は四千九十六)で漢字を紙テープ上に符号化して扱う。この符号はそのままコンピュータに入力できるものであることに着眼し、HITACのコンピュータと連動実績のある沖電気工業の機械を用いることにした。盤面に入れる漢字の種類やその排列順には、国立国語研

究所独自の方式によって仕様を作り、特別注文した。研究所はこれまでに雑誌の用語用字調査を行なっており、漢字の使用状況についてかなり精度のいい情報をもっているので、これを利用して、作業者がなるべく少ない労力で打鍵できるように字の排列を工夫した。盤面には六百のキートップがあり、各キートップが四つの文字を受けもつので、合計二千四百の字種をそなえている。片足および両足を使って4文字のシフトをする。盤内の漢字は、当用漢字から「朕」「墾」「脹」「壱」「弐」の五字を除いた千八百四十五字と、使用度数の比較的高い表外漢字二百六十四字の合計二千九字。盤面にない漢字は盤外字マークのあとに盤内ノンシフトの字を二つ続けて表わす。盤外字コード表を作って作業者があとで翻字する。例えば「迦」の字は「◆定町」と打つことになっているので、「お釈迦様」は「お釈◆定町様」と打たれてのち、もとの字に翻字される。(⇨文献6の松本論文)

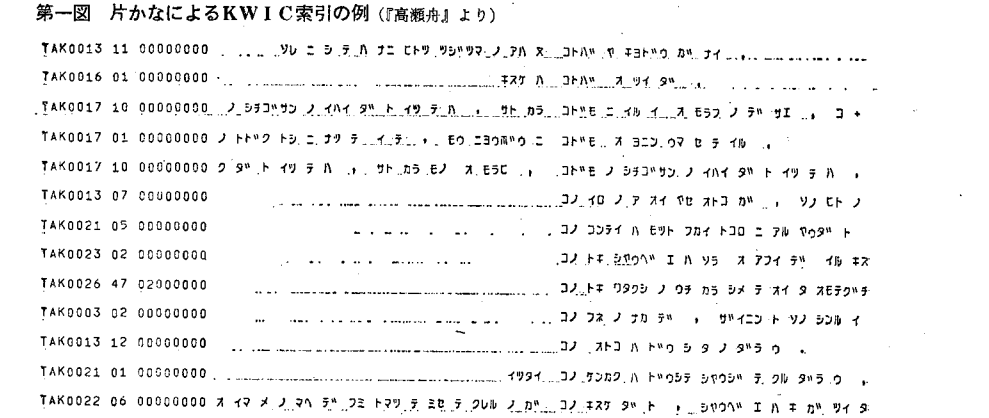
漢字をコンピュータで扱う時、一字を四けたの数字で表わすなどの方法が用いられることがあるが、あまり効率がよくないようである。漢字をコンピュータに入力するには、やはり漢字テレタイプを用いるのが最良の方法であると考えられる。この調査を企画した昭和四十年当時そうであったのはもちろんのこと、現在でもこの状況は変っていない。それに対し、漢字出力については、大分状況の変化がある。この調査では、コンピュータで処理されたデータのうち漢字テレタイプで印字しなければならない部分は、コンピュータのテープせん孔機で紙テープに出力され、その紙テープを漢字テレタイプの印字機にかけて印字した。この時の印字速度は一分間百二十字で極めて遅い。この遅さは、今回の調査における大きなネックの一つであった。現在はコンピュータからオンラインでつながる高速漢字プリンタが各社で製作されており、それらは一分間二万字から十万字以上の印字能力をもつ。印字の文字には、電光ニュースの文字のように点の集合で表わすドット方式のものと、写真植字方式のものとがある。ドット方式のものは速度は速いが字形が写真植字に著しく劣るので印刷の組版には用いがたいという欠点がある。このような高速漢字プリンタの出現によって、漢字出力の悪条件はかなり克服されてきたが、これらの機械は目下コンピュータそのものと同じくらい高価であるから、実用化にはまだ時間がかかるだろう。国立国語研究所も、コンピュータの更新には成功したが、高速漢字プリンタの導入は四十八年度には果せなかった。

【四】文脈つき用語索引(KWIC)の作成

コンピュータの補助記憶装置に単位切りのすんだ言語作品を収めておけば、比較的簡単なプログラムによって、第一図のような文脈つき用語索引を作ることができる。これは、作品中の全単語が、前後一定字数の文脈をつけたままで紙の中央に、五十音順その他指定する排列順序に並ぶものである。これは、作品に用いられた単語の全用例が所在場所の情報つきで直ちにわかるので、作品の用語を研究するのには大変便利な資料である。キーワード(Key Word)が文脈の中(In Context)にあって並ぶので、この様式の用語索引をKWIC (Key Word In Context) 索引と呼んでいる。ローマ字やかな文字で書かれた作品なら扱いが極めて容易だし、漢字を用いてあっても、高速漢字プリンタを用いれば問題はない。

国語研究所では、次のように、各種の入出力様式によるKWIC索引作成プログラムが用意してある。(⇨文献6の斎藤論文、文献8の石綿論文、文献9の土屋論文)

第一図　片かなによるKWIC索引の例（『高瀬舟』より）



〈入力機〉〈入力文字〉〈出力機〉〈出力文字〉

①フレキソ・タイプ　片かな　ライン・プリンタ　片かな

②フレキソ・タイプ　ローマ字　ライン・プリンタ　ローマ字

③漢字テレタイプ　漢字かなまじり　漢字テレタイプ　漢字かなまじり

④漢字テレタイプ　漢字かなまじり　ライン・プリンタ　片かな

①は入力も出力も能率的で速いので最も実用性があるが、文字が片かなである点にもどかしさがある。②は原文がローマ字であるものについては全く問題がないが、一般の日本文には、あまり適用の効果がない。③は漢字がそのまま生きるので日本語作品の研究資料として有用性が高いが印字速度が遅いことが大きな欠点である。④は漢字で入力されているデータをコンピュータの中で、後述の方法で片かなに自動変換してラインプリンタに打ち出すもの。すでに漢字テレタイプで入力してある資料については時間も早くてこれが最も実用的である。同じく④に属するもので、出力の片かなに漢字の漢テレ・コードを添えて、あとで漢字に翻字できるようにしたプログラムも作ってある。

現在これらのKWIC索引は、新聞用語調査のデータについて部分的に作成しているほか、夏目漱石と森鷗外のいくつかの作品について索引作成が進んでいる。漱石の『三四郎』『硝子戸の中』、鷗外の『高瀬舟』『雁』の片かなによる索引はすでにできている。古典作品についても同様の方式で索引作成を試み、江戸時代の洒落本『遊子放言』、滑稽本『浮世風呂』、古くは『今昔物語』の巻二十六と巻三十などについて作成が完了している。また、鷗外の短編『寒山拾得』については、

第二図　高速漢字プリンタで印字したKWIC索引の例（『寒山拾得』より）

思は
牧民の職に⊗て賢者を礼すると云ふのが、手柄のやうに 思は れて、⊗官絵に満足を与へるのである・ 【468・10】

思ひ
心の中では、そんな事をして⊗る寒山、拾得が文殊、普賢なら、⊗に跨つた畳干はなんだらうなどと、田舎者が芝居を見て、どの役がどの俳優かと 思ひ 惑ふ時のやうな気分になつて⊗るのである・ 【472・11】

思ふ
それから二人顔を見合せて腹の底から篤み上げて来るやうな笑声を出したかと 思ふ と、一しょに立ち上がつて、⊗世係を駆け出して逃げた・ 【474・4】

帰つ
或る日山から⊗に騎つて 帰つ て参られたのでございます・ 【470・4】

返つ
僧は振り 返つ た・ 【465・11】

帰ら
それは先頃まで、本堂の背後の僧院にをられましたが、行脚に出られた切、 帰ら れませぬ・ 【469・9】
あれは畳干さんが松林の中から拾つて 帰ら れた捨子でございます・ 【471・10】

願
自分の職業に気を取られて、唯営⊗役⊗と年月を送つて⊗る人は、道と云ふものを 願 みない・ 【467・3】

漢字テレタイプで入力したデータをコンピュータでコード変換し、高速漢字プリンタ GEM 3800 にかけて印字することを試みた。第二図がその一部である。

B 言語処理自動化の研究と実験

新聞用語用字調査の実際処理の中には、自動処理といえる要素は、ほとんど無い。全体処理の、初め、中、終りに人間が介入して、いろいろな処置を施している。しかし、コンピュータを使うのは、もともと省力化のためであるから、将来の言語調査においては、人力作業の介入をできるだけ減らしたい。そこで、新聞用語用字調査を進めるかたわら、その調査資料を用いて、処理自動化の実験をさまざま試みた。それには、次のようなものがある。

【一】単位切りの自動化

前述のとおり、新聞用語調査では、長短二種類の単位を設けて、文を単語に分割した。作業はすべて人力でなされ、これに多大の労力と時間を要した。この作業を自動化することができないであろうか。われわれの知っている大ていの外国語では、文章を書くとき、単語と単語の間をあけて書く習慣があるから、一見して単語の区別が明らかである。日本語の漢字かなまじり文では、句読点のある処以外は、文字が連続して書かれるから、コンピュータが無条件に単語の切れ目を認定することはできない。そこで、単位切りにどうしても人力が必要になる。機械翻訳にしても、情報検索にしても、日本語の文章を扱うところには、まず単位切りの作業が必要になる。この作業の自動化ができれば、日本文のデータ処理に革命をもらすことになろう。

日本語における単語認定は、新聞用語調査で長短二種類の単位を認めなければならなかったことにも表われているように、いろいろな考え方が可能なので大量のデータについて完全に一貫した原理で単位切りを果すことは、人間にとっても、非常にむずかしい。コンピュータに単位切りをさせるのは非常にむずかしい課題だが、ひとたび成功すれば、人間と違って、判定基準を狂わすことがないから、どんなに大量の処理をしても作業の一貫性がくずれるおそれはなくなる。

国立国語研究所では、次の二つの考え方で、この問題に迫った。

1 文字排列の情報を利用する方法

漢字かなまじり文においては、助詞・助動詞などの付属語はひらがなで書かれ、名詞・動詞・形容詞・形容動詞などの自立語は、漢字で書かれる傾向がある。この事実に着眼して、漢字かなまじり文について、漢字とかなとの間で切断すると、

春一が一来一た一。

今回一の一事件一の一黒幕一は一君一らしい一。

のように、それだけで、単位切りが成功する場合がある。いつもこううまくはいかないが、この原則を破るものには、①動詞・形容詞は、語幹を漢字で、語尾をかなで書くことが多い。②動詞・形容詞でも「いう」「いく」「ある」「する」「いい」「ない」のような語は、かな書きになりがちである。③副詞は、かなで書かれることが多いが、「最も」「優に」のように漢字とかなで書かれる語も少なくない。——など、いくつかの規則的な条件が見出される。動詞・形容詞の語尾は形に制限があるし、後続するかなとの間に一定の関係が見出される。かなで書かれやすい自立語は特定なもので、数が多くないから、一覧表を作って参照させればよい。こういう処理過程を判断回路に組んで、前述の原則に加えると自動単位切りプログラムを作ることができる。小実験では、ある程度の成功を見たが、大量

のデータについて実用化を図るには、まだ多くの研究を要する。(↓文献13の江川論文)

2　辞書引きと構文解析による方法

右に述べた方法は標準的な漢字かなまじり文にでなければ適用することができない。漢字の極度に少ない文や、かなだけの文に対しては、単位切りの手がかりがないことになる。もっとも、普通、かなだけで文章を書く場合は、電報のようにのべつに書いては、人間にも読みにくいので、初めから単語の間をあけて書く(これを「分かち書き」という)ことになるのが普通だが、そうすると、今度は分かち書きの原理が一貫しなくなるおそれがある。いずれにしても、漢字かなまじり文はもちろんのこと、たとい、かなだけで分かち書きのしてない文章でも、正しく単語に切ることのできる方法を求めておかなければならない。この必要から考えられたのが、辞書を引きつつ単語を認定していく方法である。

例えば「はながさく。」という文がある場合、まず「は」の一字を取って、記憶装置内の辞書と照合し、一致するものがあれば、ひとまず、「は」を単語と認定する。次に「はな」についても同様の処置をする。以下、「はなが」「はながさ」「はなざさく」について、順次同様に辞書と照合した結果、「は」「はな」「はながさ」の三種が単語の可能性をもっと認定されたとする。「は」を単語としたら、次は「な」「ながさ」「ながさく」を辞書と照合して、「な」と「ながさ」が単語と認められたとする。「は+な」のあとは「が」「がさ」「がさく」について辞書と照合し、「が」だけが単語とされる。「は+な+が」のあとは、「さ」「さく」を辞書と照合し、「さく」が単語とされる。これでひとまず「は+な+が+さく」のつながりが得られる。次に、「は+ながさ」のあとは「く」だけが残る。「く」が辞書照合で単語と認められれば「は+ながさ+く」のつながりが得られ、もし「く」が認められなければ、「は+ながさ」のつながりは棄てられる。このようにして、順次、可能性をさぐって行って、結局「は+な+が+さく」「はな+が+さく」「はながさ+く」の四種類に単語のつながりの可能性が認められたとする。ここで止まれば、でたらめな答えを含んでいることになるので、次に、辞書に添えてある品詞情報を参照して、

①「は+な+が+さく」を「名詞+名詞+助詞+動詞」に

②「は+ながさ+く」を「名詞+名詞+名詞」に

③「はな+が+さく」を「名詞+助詞+動詞」に

④「はながさ+く」を「名詞+名詞」にそれぞれ変換する。この四種の品詞構成を、貯えてある構文規則と照合していくと、③の場合だけが文の形をしていると認められる。あとは構文規則に合わないので文をなしていないと認められ、その仮設が棄てられる。こうして、辞書引きののち、構文解析による検定を経て、「はな」が「さく」という単位切りが完了する。

この方法が功を奏するか否かは、専ら、辞書と文法規則が完備しているかどうかにかかっている。やたらにたくさん辞書引きがあって能率が悪そうな感じもするが、こういう単純なことの繰り返しは、コンピュータにとっては何でもないことである。それも、補助記憶装置が磁気テープだけだと照合に時間がかかるが、磁気ディスクや磁気ドラムのようなランダム・アクセス装置があれば、ずっと能率がよくなる。

ここでやっかいなのは、例えば「は」という字は、「ハ」とも読めるし「ワ」とも読め

て、どちらに読んだかによって単語認定の結果が変って来る、というようなことが起ることである。かなで二様に読めるのは「は」と「へ」だけだが、漢字には読みかたがいろいろある場合が少なくない。例えば「目下」は「もっか」とも「めした」とも読め、「もっか」なら副詞、「めした」なら名詞と認められる。このようなあいまいさがある場合は、一つの処置だけで次に進まないで、あらゆる可能性をたどって、いくつもの解答を出すように周到なプログラムを作る必要がある。

この方法は、極めて小量の文章について試みてみただけであり、実用化はまだ遠いが、可能性は充分にある。(⇩文献7の木村論文、文献12の石綿・斎藤・木村論文)

【二】よみがなつけの自動化

「目下」のように両様に読める例は文章の中にそう多く出て来るものではない。例えば「今は花が最高に多い季節だ。」という文は、だれが読んでも「イマはハナがサイコウにオオいキセツだ。」としか読めない。ここに迷いがないのには、二つの原理がひそんでいると思われる。その一つは、われわれが、ここに用いられている単語をよく知っているということだ。「今」「花」「最高」「季節」を、一見して直ちに単語と認めるから、その瞬間に「イマ」「ハナ」「サイコウ」「キセツ」の読みがきまるのである。「今」は「イマ」とも「コン」とも読めることは知っているが、「今」で一語と認めた時に「コン」は棄てられる。「最」「高」がそれぞれ一語なら「モットモ」「タカ」と読まれるが、この場合は「最高」で一語だと認めたから、「モットモ」も「タカ」も出て来る余地がなく、直ちに「サイコウ」と読まれたわけである。もう一つの原理は、われわれが漢字に反応する時に「音」反応の系と「訓」反応の系をもっていて、漢字一字に対しては、とかく「訓」反応の系が働き、漢字二字の連続には、とかく「音」反応の系が働くということである。これら二つの原理をそのままコンピュータのプログラムにもちこんで、漢字解読の自動化を図ってみた。ここでいう二つの原理は、結局、前項の自動単位切りの二つの原理と同じことで、前者は辞書引きによる法であり、後者は文字排列の情報を利用する法である。

1　文字排列の情報を利用する方法

漢字がなまじり文において、平がなに囲まれた漢字が、一字だけなら、それに訓を与え、二字以上なら各字に音を与えることを基本原則とし、漢字の音訓表を作っておいて、文中の各漢字に周囲の文字環境によって音か訓かを選ばせるようにする。この原則には、重箱、湯桶読みを始め、例外がいくらでもある。その例外を辞書に答えさせるのでは、問題が字のレベルから語のレベルへ移ってしまう。つまり、文が単語に切れていて、かつ大きな辞書をそなえていなければならないことになる。それを避け、例外も、文字排列情報を利用し、簡単な環境演算回路と簡単な音訓表とによって処理してみたい。事実、それはある程度可能なのである。例えば、「崩」という字は「崩御」「崩壊」のように後続字が漢字の場合は「ホウ」と音読されるが、後続字がかなら、先行字が漢字でもかなでも「山崩れ」「…が崩れる」のように「くず」と訓で読まれる。「板」は、前後とも漢字、または前だけが漢字の時に「看板娘」「黒板ふき」のように「バン」と音読され、それ以外の時は「板前さん」「杉の板を切る」のように「いた」と訓読される。このように、漢字によっていくつかのタイプがあるので、それに対応できるように音訓表を用意し、環境演算の回路を、その字の特徴によって指定できる

ようにしておけばよい。第三図は、このようにしてコンピュータが漢字を読んだ結果の一例である。「引続き」や「見通し」が「いんぞくき」「けんつうし」と読まれてしまっているが、全体としては86%程度の正解を得た。(⇩文献7の田中論文) 単位切りも辞書も要らないのがこのプログラムの特長である。

第三図　自動よみがなつけの例〈その1〉(昭41 朝日新聞紙面より)

ベトナム問〔もん〕題〔だい〕をめぐる米〔べい〕国〔こく〕の〝平〔へい〕和〔わ〕外〔がい〕交〔こう〕攻〔こう〕勢〔せい〕〟は，年〔とし〕が明〔あ〕けても引〔いん〕続〔ぞく〕き行〔ゆ〕われ，北〔ほく〕爆〔ばく〕停〔てい〕止〔し〕も続〔ぞく行〔こう〕している．しかし問〔もん〕題〔だい〕のカギを握〔にぎ〕る北〔きた〕ベトナムは，これまでのところ従〔じゅう〕来〔らい〕通〔とお〕りのきびしい反〔はん〕応〔のう〕しか示〔しめ〕しておらず，ワシントンでは悲〔ひ〕観〔かん〕的〔てき〕な見〔けん〕通〔つう〕しも現〔あらわ〕れ始〔はじ〕めた．ドゴール仏〔ふつ〕

2　辞書引きによる方法

単位の切ってある文章についてならば、辞書の照合によって、漢字で書かれた単語によみがなを与えることは容易である。まず辞書のファイルを作ることが先決問題だが、新聞の用語調査で語彙表ができているので、それに多少の情報を加えることによって、約十万語の辞書ファイルを作った。この辞書は、今後見出し語の数をふやしつつ、いろいろな情報をつけ加えていけば、各種の言語処理に役立つ汎用辞書ができる。今は漢字によみがなをつけることだけが目的だから、よみがなの情報さえあればよい。入力した文章の中に、辞書に収録されてない語があれば、まとめてリストにして印字し、あとで人間がよみがなを与える。以後、その語は辞書に増補されることになる。このプログラムを二、三の記事でテストした結果99%に近い正解率を得た。第四図はその一部である。残りの1%の誤答は、「中〔ちゅう〕佐〔さ〕は驚〔おどろ〕いた風〔かぜ〕である。」「雨〔あめ〕が降〔お〕り」「仏〔ほとけ〕がNATOを脱〔だっ〕退〔たい〕」「高〔こう〕校〔こう〕に進〔しん〕学〔がく〕される方〔ほう〕のために」などであった。

第四図　自動よみがなつけの例〈その2〉
(昭41 毎日新聞紙面より)

K001129010101501 ＃ 大〔おお〕谷〔たに〕 重〔じゅう〕工〔こう〕 深〔ふか〕川〔がわ〕 工〔こう〕場〔じょう〕 ポヤ ＃

K001129010401050 2 ＃ 三〔みっ〕日〔か〕 午〔ご〕後〔ご〕 一〔いち〕時〔じ〕 十〔じゅう〕五〔ご〕分〔ふん〕ごろ， 東〔とう〕京〔きょう〕 墨〔すみ〕田〔だ〕 区〔く〕 柳〔やな〕原〔はら〕 町〔ちょう〕 一〔いち〕 ○〔れい〕 ， 大〔おお〕谷〔たに〕 重〔じゅう〕工〔こう〕 深〔ふか〕川〔がわ〕 工〔こう〕場〔じょう〕 の 機〔き〕械〔かい〕 工〔こう〕場〔じょう〕 南〔みなみ〕側〔がわ〕 の カベ から 煙〔けむり〕 が 出〔で〕ている のを 従〔じゅう〕業〔ぎょう〕員〔いん〕 が みつけた。＃

3　活用形処理の自動化

動詞、形容詞、形容動詞および助動詞は、語尾が、「か、き、く、け、け」だの「く、い、い、けれ」「だろ、だっ、に、で、だ、なら」などと変化する。人間なら、語尾が変化しても同じ語だと認めて、例えば、「書か2」「書き3」「書い4」「書く1」「書け1」「書こ2」とあれば、これら全部を集めて「書く13」とする。このように、各活用形に散らばっているものを一つの代表形のもとに集める作業をコンピュータにさせるプログラムを作ってテストした。結果はかなり好成績であった。時間をかけて諸種の参照表を充実させ(⇩文献8の石綿論文)

【三】その他の試み

以下は、くわしい説明を省き、項目をかかげて内容の概略だけを記す。

1　かなから漢字への自動変換

漢字によみがなをつけるのは漢字をかなに変換することであるが、その反対に、かなを漢字に変換することも必要になる。コンピュータにかなだけの文章を入力し、それを漢字かなまじり文にして出力させることを試みたが、これは、漢字からかなへの変換よりもむずかしい問題である。(⇩文献14の田中論文)

2　活用形処理の自動化

動詞、形容詞、形容動詞および助動詞は、

れば、実用化の可能性が充分ある。(⇩文献7の江川論文)

3　品詞認定の自動化

単位切りがなされている文章について、各単語の品詞をコンピュータに判定させようとするプログラムで、①辞書を引いて決める方法②その語の語形から判定する方法③直前の語からの接続のしかたで判定する方法の三つの方法で小実験を試みた結果、それぞれ一長一短があるが、正解率は80%から93%の間であった。(⇩文献8の中野論文)

4　同語異語判別の自動化

これはまだ、試みたとはいえないが問題点だけ述べておこう。コンピュータで用語調査をした結果「いっ」という語形が10箇拾われたとする。この中には、「言っ」と解さるべきものも「行っ」と解さるべきものも、いっしょになっているから、ことばの調査として正確を期するためには文脈を調べて「い(言)っ6」「い(行)っ4」のように区別しなければならない。そして「い(言)っ」の方は「言わ」や「言い」といっしょに「言う」の中に入れ、「い(行)っ」の方は「行か」や「行き」といっしょに「行く」の中に入れることによって、はじめて正しい用語調査となる。「いっ」を「い(言)っ」と「い(行)っ」とに分けるのは同形異語を判別する作業、「い(言)っ」「言っ」「言わ」「言う」「いう」などをいっしょに扱うのは異形同語を判定する作業である。

同形異語とは、別々の単語がたまたま同じ字で書かれることで、文脈から切り離してそれだけ見た時には区別がつかない。「いっ」は、かなによる同形異語の例、「工夫(こうふ・くふう)」「方(ほう・かた)」などは漢字による同形異語の例である。異形同語は、本来同一の語が文中でいろいろな形をとって現われるもので、この現象は、活用による語形変化と、漢字・かなによる表記法のちがいとから生ずる。同形異語の判別と異形同語の判別とをいっしょにして同語異語の判別といっている。同語異語判別の自動化ができないと、用語用字調査の完全自動化はできないし、機械翻訳以下言語自動処理のすべてにわたって障害が残る。同語異語の自動判別は、日本語のコンピュータ処理の足元にある重大な課題である。

C　言語処理自動化のための基礎的言語研究

言語処理が行きづまるところには、必らず言語研究の基本問題が横たわっている。言語研究のためにある国立国語研究所にとって、そういう基本問題の研究は、当面の処理業務におとらず大切である。研究所がこれまでに行なってきた研究は、どの部門の研究も、みな究極においては言語処理法の開発に道を開く研究として意味をもつが、ここでは、直接、言語処理自動化を目ざしつつ、かつ、現実にコンピュータで作成した研究データを使って行なった研究を三つ取り上げて紹介しておこう。

【一】日本語の動詞句形成パタンの研究

所員石綿敏雄は、日本語文の自動構文解析の文法規則を精密化することを目的として、「名詞+助詞+動詞」の形で作られる動詞句における語の意味類型の関係を分析し、格助詞「に」や「を」を含む動詞句について単語排列のパタンを記述した。この種の研究には大量のデータが必要なので、新聞用語調査や『三四郎』から KWIC 索引を作り、多数の用例を分析した。これによって見出された動詞句「名詞+に+動詞」の形成パタンは、概略次のようなものである。

一一一二つの名詞を特定の関係でつなぐ動詞句

〔名詞一般〕**に**〔抽象関係を示す動詞〕

（例）国際経済に関する報告、諸国に対する援助

二―一 移動を表わす動詞句
〔位置を表わす名詞〕に〔移動を表わす動詞〕
（例）病院に行く、左右に動かす

二―二 意図的行動を表わす動詞句
〔人間行為を表わす名詞〕に〔移動を表わす動詞〕
（例）手術に行く、見に来る

三―一 存在または空間性の動作を表わす動詞句
〔場所・方向を表わす名詞〕に〔存在・動作を表わす動詞〕
（例）公園にある森林・顔を横に向ける

三―二 対象物のある動作を表わす動詞句
〔具体物を表わす名詞〕に〔動作を表わす動詞〕
（例）手に持つ、電車に乗る

四 事態の発展・推移を表わす動詞句
〔時間概念を含む名詞〕に〔抽象的意味の動詞〕
（例）将来に備える、遠足が三日にのびた

五 人間的存在に対する行為を表わす動詞句
〔人間や人間集団を表わす名詞〕に〔行為を表わす動詞〕
（例）親に甘える、仲間に知らせる

六 句全体で一つの行為を表わす動詞句
〔人間の行為を表わす名詞〕に〔動詞一般〕
（例）相談に応ずる、失敗に終る

この研究において、名詞や動詞の意味の分類には国立国語研究所編『分類語彙表』（昭39）を使っている。（⇨文献9の石綿論文）

【二】漢字かなまじり文における文字連続の研究

所員野村雅昭は、新聞記事を材料に、漢字かなまじり文を文節に切り、各文節内における漢字かなの排列パタンを分析して、次のような傾向を見出した。

(1)文節は、漢字で始まり、ひらがなで終る傾向が強い。
(2)ひらがなで始まる文節のうち、接続詞・連体詞・副詞などは、ひらがなだけの文節を構成しやすい。
(3)ひらがなで始まる文節のうち、名詞・動詞・形容詞は、直前に、ある範囲内のひらがなの来る確率が高い。
(4)ひらがなで書かれる名詞・動詞の大部分は「こと」「もの」「する」「いう」など、使用頻度の高いものである。

（⇨文献9の野村論文）

【三】指示語「この」「その」の意味の受けつぎ方の研究

本稿の筆者林は、鷗外『高瀬舟』のKWIC索引を材料にして「この」「その」の用法を分析し、それらの指示機能と前後のことばの形式上の特徴に次のような傾向を見出した。

(1)「この」「その」が現場の存在物を指示する時は、先行文脈中に指示目標がない。
(2)「この」「その」が先行文脈中の語を代行する時は、「この」「その」の被修飾語に関係概念や、全体に対する部分を表わす語が来やすい。
(3)「この」「その」が文脈中の語を指定する時は、近い先行文脈中にそれらの被修飾語と同形または極めて近縁の語があることが多い。
(4)「この」「その」が文脈中の語を指定していても、「この」「その」の被修飾語が時間条件を表わす形式名詞の場合には、(3)の条件をもたない。（⇨文献9の林論文）

以上、国立国語研究所におけるコンピュータ言語学の概要を述べた。これらの研究は始まったばかりで、まだ幼稚な段階にある。目を外国といわず、日本の各地に向けるだけで

も、東京大学情報科学研究施設、同音声医学研究施設、京都大学情報工学科、東北大学電気通信研究所、九州大学通信工学科、神戸大学経済経営研究所、通産省電子技術総合研究所、日本科学技術情報センター、国立国会図書館、日立中央研究所、その他、多くの大学や研究機関で、漢字入出力、音声の分析・合成、機械翻訳、情報検索、自動抄録、マン・マシン・コミュニケーション、等の業務開発を通じ、音声、音韻、語彙、意味、文法、文章構造等の基礎的言語研究が堅実に行なわれていることを知って驚く次第である。今後これらの研究機関と情報を交換しながら、研究所独自の問題把握で、コンピュータ言語学の課題に一つ一つ取り組んでいきたい。

参照文献（1から10までの文献はすべて国立国語研究所の著作で秀英出版から刊行されている）

1『電子計算機による新聞の語彙調査』昭45
2『電子計算機による新聞の語彙調査Ⅱ』昭46
3『電子計算機による新聞の語彙調査Ⅲ』昭47
4『電子計算機による新聞の語彙調査Ⅳ』昭48
5『現代新聞の漢字調査（中間報告）』昭46
6『電子計算機による国語研究』昭43（林四郎「新聞語彙調査の概略と語彙分析法試案」石綿敏雄「語彙調査第一段階のプログラムの基本的な考え方」「言語の意味と言語情報処理」松本昭「国研用漢字テレタイプと同機利用の言語情報処理」斎藤秀紀「電子計算機と漢テレによる用語総索引の作成」田中・斎藤「新聞語彙調査のサンプリング・プログラム」田中章夫「電子計算機によるワードリスト作成上の一問題」木村繁「漢テレ入力データのチェック」）
7『電子計算機による国語研究Ⅱ』昭44（林「新聞語彙調査における層別とその意味」斎藤「電子計算機による語彙調査」木村「層別特徴語の判別」「構文解析自動化の研究Ⅱ」中野洋「語彙調査の類別語彙表について」江川清「活用形処理の自動化に関する一方式」石綿「COBOLによる漢字索引作成」「構文解析自動化の研究」野村雅昭「新聞使用漢字の試行的分析」田中「漢字かなまじり文を全文カナ書き・ローマ字書きに変換するシステムについて」）
8『電子計算機による国語研究Ⅲ』昭46（林「語彙調査と基本調査」石綿「新聞用語調査の用例印字プログラム“COBOL-KWIC”」斎藤「電子計算機による語彙調査Ⅱ」中野「品詞認定の自動化」田中「新聞語彙調査の同音語と同形語」野村「新聞漢字調査の機械処理システム」）
9『電子計算機による国語研究Ⅳ』昭47（野村「漢字かなまじり文の文字連続」土屋信一「カナ入力による日本語文総索引の作成」村木新次郎「あいまいさを伴う表現の構造についての一考察」石綿「助詞『に』を含む動詞句の構造」林「指示連体詞『この』『その』の働きと前後関係」）
10『電子計算機による国語研究Ⅴ』昭48（石綿「電子計算機による語彙調査と同語異語の処理」「人間の精神活動を意味する動詞の用法」斎藤「電子計算機による語彙調査Ⅲ」村木「用語の集中度と共通度」野村「複次結合語の構造」中野「現代日本語における音素連続の実態」鷲岡昭夫「文語形・口語形活用語の代表形の変換処理について」田中「自動抄録処理におけるキー・ワードの性格」林「コンピュータによる言語資料の研究（英文）」）
11田中章夫「漢字の自動解読システムについて」計量国語学48号、昭44
12石綿・斎藤・木村「言語単位分割自動化の研究」計量国語学50号、昭44
13江川清「単位分割自動化のシステムについて」計量国語学51号、昭44
14田中章夫「ヨミガナ方式によるカナ（ローマ字）の漢字変換」計量国語学55号、昭45

（林　四郎）