計算言語学の立場からの提言

田中　穂積（東京工業大学）

【研究の道具としてのコンピュータ】

これまでコンピュータは高価で一部の研究者が独占的に使用する道具であった。しかし、最近のコンピュータ技術の進展ぶりは目ざましく、コンピュータの著しい性能向上とともに、以前とは比較にならないほど安価になってきている。研究者にとってコンピュータは身近な存在になってきている。コンピュータの進展は社会だけでなく科学・工学分野の研究に大きな影響を与えてきた。これまで、科学・工学の分野、特に科学計算の分野では、コンピュータは研究を効率よく進めるための道具として大きな役割を果たしてきた。言語学についてもコンピュータを道具として利用することができる。一般に研究の道具としてコンピュータが果たす役割をまとめると次のようになる。

コンピュータの内部に理論や現実の世界をプログラムとしてモデル化し、プログラムを動作させてプログラムの挙動を調べることにより、理論の妥当性を検証したり、コンピュータ内部に実現したモデルの妥当性を調べることができる。プログラムによりコンピュータは検証しようとする理論や現実世界のシミュレータに変身する。このシミュレータの特徴は細部にわたる動作の追跡が可能なこと、繰り返し実験が可能なことにある。しかもプログラムはパラメータの設定を変えたりプログラムそれ自身を修正することが容易であるため、コンピュータ内部に様々なモデルを作り出すことができる。

この様にコンピュータは従来の道具（実験装置）と比べて格段に柔軟な実験装置であるといえる。最近のハードウェア技術の目ざましい進展によりコンピュータの高速化が進み、実験時間も大幅に短縮できる。実験時間の短縮は研究の効率性と進展とに大きく寄与する。

ここでプログラムの性質について一言述べておきたい。プログラムはコンピュータと一体になり、様々な種類の問題解決を行うマシンに変身することができる。その意味でプログラムは問題解決を行うマシンであると言ってもよい。プログラムが高水準のプログラム言語で書かれていれば、コンピュータの種類を問わず実行することができる。コンピュータが変わってもプログラムはそのまま生き続けることができるのである。これはこれまでのマシンにない特徴で

ある。

【言語とコンピュータ】

言語学の研究にコンピュータを利用することは、国立国語研究所を始めとして既に多くの実績がある。コンピュータの本質が記号を操作するマシンであることは良く知られている。一方言語は記号の系列であり、言語を使用することは記号操作であるから、コンピュータは言語を使用することができる最初のマシンであるといえよう。言語とコンピュータは、コンピュータの作りからして相性が良いと結論することができる。言語とコンピュータを巡る問題を思い付くままに列挙してみよう。

(1) 大量の言語データの蓄積と統計的な分析を行う。

(2) 言語学の成果をコンピュータにのせて、

(a) 言語理論の妥当性を検証する、

(b) 言語理論を精密化する、

(c) 言語理論を修正する。

(3) 言語処理用のアルゴリズムを開発する。

(4) 言語理解過程のモデル化を行う。

(5) 言語理解システムを作成する。

以上は計算言語学の典型的な研究テーマである。機械翻訳システムや、自然言語理解システムの開発は(5)に含まれる。

【言語データの蓄積】

国立国語研究所でコンピュータを利用した言語学の研究が行われてきたということを既に述べた。研究の多くは主として(1)であったように思う。(1)に関連して(3)の研究も行われている。

言語の研究で大量のカードを用いることが良く行われている。確かにカードに書き留めた情報は、コンピュータの使えない電車の中や自宅でも取り出して見ることができる。そういう利点はこれまで確かにあった。しかしカードの容量が増えるにつれて検索に要す時間も増える。持ち運びも不便になろう。これに対して、最近のコンピュータ技術の進展により、研究所におかれたコンピュータ上のデータを、自宅や電車の中でも簡単に取り出せるような時代が到来し

ようとしている。現在の情報検索技術には研究の余地があるにしても、コンピュータによる情報検索は人手を大幅に軽減し高速化することができる。

コンピュータの記憶容量は最近飛躍的に増えている。価格も安価であり、昨年の筆者の経験では、150万円で680ＭＢの二次記憶装置が購入できた。予算的な障害は以前ほどでないと思われる。むしろコンピュータを利用した言語データの蓄積に大量の労力が必要であることが問題である。確かイタリアのピサ大学であったと思う。今から4年ほど前に、中世の文学作品を全てコンピュータに入力し電子化する計画を進めていた。中世の日本文学作品を電子化する計画は、筆者が以前所属していた通産省ではできない。国立国語研究所で是非進めてほしい。それとともに、言語の研究を進めるためにあらゆる言語データを国立国語研究所で蓄積し、研究用として公開して欲しいものである。

言語データの蓄積に際し次の注意が必要であろう。言語データとして加工したものを蓄積することは好ましくない。蓄積する言語データはよほどの理由がないかぎり生データに限るべきである。加工に際し意見の一致が難しいことがしばしばあるからである。どのような加工を施すかは生データを使用する研究者に任せるべきである。研究者はコンピュータの力を借りて比較的容易に加工を施すことができる。加工についての議論は、未加工の生データを蓄積してからでも遅くない。

このことに関連し、当時電子技術総合研究所の淵一博氏（現新世代コンピュータ技術開発機構研究所長）が電子化した新明解国語辞典が思い起こされる。必要以上の加工を避けて電子化されているので校正作業が容易になり、その後新明解国語辞典を利用した研究が各所で行われている。

【用例ＫＷＩＣと国語辞典】

国立国語研究所の重要な任務の一つが国語辞典の作成にあることは良く知られている。国立国語研究所で新聞データの統計的な分析が行われたのも、初期の段階ではこの任務が意識されていたからだと思われる。たとえば語の使用頻度は語の重要性の一つの指標になる。しかしそれ以上に重要なことは、特定の語を深く分析することだろう。そのために言語データに関するＫＷＩＣが有効である。ＫＷＩＣを利用すれば語の用例をたちどころに収集することができるからである。

ところがＫＷＩＣがＫＷＩＣとして有効になるのは、大量の言語データを対象とした場合であるといわれている。少量の言語データのＫＷＩＣでは用例不足に悩まされ役に立たない。かって第５世代コンピュータ計画で百万例文のＫＷＩＣを開発する構想があったが、通産省の研究テーマとして成立しにくいとの理由で実現しなかった。第５世代コンピュータ計画での構想は、その後電子化辞書計画として通産省の肝入りで実現した。電子化辞書計画は将来の自然言語理解システムや機械翻訳システムに必要とされる machine readable な辞書を7年計画で開発するものである。電子化辞書計画は民間の電子化辞書研究所（ＥＤＲ）で推進されているが、その計画によれば政府資金が70％で総額60億円ほどが予定されている。

最近ＥＤＲを見学したおりに、一千万例文のＫＷＩＣを作成する計画があることを知らされた。日本語のＫＷＩＣを作成するためには、例文に対して形態素解析を施す必要がある。ＥＤＲでは2台のワークステーションをフル稼働させ、日本語の例文を形態素解析し、分かち書きすると共に、形態素解析の過程で推定された未定義語の抽出作業を既に進めている。

このような膨大な例文ＫＷＩＣは、国語研究所で蓄積すべきものであったと思う。既に述べたように国語辞典の開発には用例の収集と分析が不可欠であると思われるからである。にもかかわらず国語研究所で大規模な例文ＫＷＩＣが作成されなかった理由はいくつか考えられる。予算面の制約から、膨大な量の例文ＫＷＩＣの計画が実現不可能に追い込まれたのかも知れない。計画立案段階ではコンピュータのパワー不足もあったのかも知れない。単なるＫＷＩＣ作成自体は研究課題になりにくいことも問題であったかも知れない。研究者がこのような力ずくの作業に積極的に取り組みたくないこともあったかも知れない。

しかし国語辞典の実現に向けて、大量の言語データの蓄積、その自動分かち書き、そしてＫＷＩＣ作成、用例の収集と分析作業という過程は、長いようで実は一番の近道であると思われる。語の出現頻度の調査は分かち書きされた言語データがあれば直ちにできる。その意味で軽量級の作業であるといえよう。それに対して例文ＫＷＩＣの作成は重量級の作業である。重量級の作業の結果を生かす研究が今後必要だと思われる。それには研究予算が相当額必要であるという主張をこれから繰り返し行う必要があろう。

ＥＤＲは民間会社であるとはいえ政府の資金が入っている。ＥＤＲのＫＷＩ

Cが近く部分的に完成し公開されることを筆者は願っている。このＫＷＩＣは国語辞典だけでなく、将来の日本語研究に計り知れない影響を及ぼす可能性がある。言語学者の研究は例文の分析から始まるからである。その意味でＥＤＲで作成中のＫＷＩＣに対して、国語研究所として利用の道を開いておくことが必要かも知れない。

【計算言語学と国語研究所】

計算言語学での典型的な五つの研究課題を挙げておいた。国語研究所ではこれまで(2)、(4)、(5)の研究はほとんど行われていない。最近の計算言語学の進展を見ると、筆者は計算言語学の国立研究所がわが国にもあったらと思う。国語研究所はこれらの研究分野を扱うことが可能な立場にある。欧米での言語学者に数学科出身者が多いと聞く。言語学は自然言語の秩序を研究するものである。秩序を形式と言い替えれば、言語の研究に理学と工学からの貢献がこれから求められるだろう。そのような人材確保が国語研究所として今後必要であると思われる。(2)、(4)、(5)に挙げた研究課題は国語研究所にとって無関係のように見えるかも知れない。しかし、日本語の基礎的な研究を幅広く行うために、これらの研究課題を取り上げることは、少なからぬ貢献をなすと思われる。

これまでの言語学は統語論が中心であった。しかし(4)と(5)の研究を進める過程で、あまりかえりみられなかった一文を越えた談話構造や照応関係の分析、トピックや焦点の本質、言語理解に関わる知識と推論、省略の問題等の重要性が認識されている。これらは語用論と呼ばれることがある。筆者は統語論の重要性を否定するものではない。統語論は統語論として研究されねばならない。しかし、一文中心の統語論に決着がついてから、語用論の問題を解決するというより、両者を並行して研究すべきだと思う。語用論の問題をこれまで言語学者が避けてきた理由は、統語論に比べて語用論が困難であると考えたからであるが、果してそう言いきれるだろうか。

【国立の研究所の役割】

筆者もかって通産省の研究所に所属していたことがあるので、国立の研究所の果たすべき役割について述べてみたい。国立の研究所では、金額的には大学

ではできず、企業の研究所ではリスクが高いとして手を付け難い研究を行うべきであるとする考えがある。このような考え方では、膨大な資金を要す核融合等のビッグプロジェクトを除き、国立の研究所が研究所として生き抜く道は狭いものになる。大学でも以前とは比較にならないほどの予算を獲得する道が開かれてきている。さらに民間の研究所でも、リスクの高い研究に着手するようになってきている。両者は国立の研究所のカバーする範囲を両面から侵食し始めている。

筆者は国立の研究所は大学でも企業でもできない研究を行うべきだという考えには賛成しない。大学でも企業でもできる研究であっても、国立の研究所が必要だと思う研究は積極的に取り上げて欲しいと思う。プロの研究者のいる国立の研究所の研究成果は、アマ研究者集団である大学とは自ずと異なる成果が得られるだろう。企業は企業の論理があり、研究の進め方も国立研究所とは異なるだろうから異なる成果が得られるだろう。この時忘れてならないものは、国立研究所から生み出される研究成果は、その分野に新しいパラダイムを切り開き、そして息の長いものでなければならないということである。

最後に日本語の研究には日本語以外の言語の研究も必要であることを指摘しておきたい。その意味で筆者は国語研究所でなく言語研究所という名前の方が好きである。日本語を中心としてあらゆる言語の研究を行ったら良いと思う。計算言語学も行うべきだろう。名前にこだわる必要はないとする考えもあるかもしれない。しかし新しい展開を図ろうとしたとき、名前が障害になることもある。筆者の所属していた通産省の電子技術総合研究所はかって電気試験所と呼ばれていた。言葉を大切にしない国は国としてのアイデンティティを失うことになる。国立国語研究所の責任は大きいものがある。今後の一層の発展を期待したい。