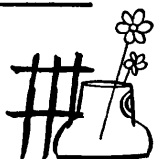


## 巻頭書



## 日本語と機械

坂井 利之†



近年日本語ブームである。出版界では、日本語関係の本が人気を呼び、コンピュータ業界では、漢字処理の各種製品を次々発表している。学会では、コンピュータ向きの日本語辞書、日本語処理など人々の注目をひき、研究も活発で誠に喜ばしいことである。これらは、日本語は漢字かな混り文であって、すべての点でアルファベット系の言語に比較して、機械との結合に對し著しいハンデキャップを持っているという固定観念打破に対する努力であると言ってよからう。

機械による言語処理で、表層に現われた日本語文章について論議するのは第一歩である。さらに言語のもっている構造そのもの、その言語を使用する人々の発想や脳内における処理と記憶の仕方、社会環境での言語のもつコミュニケーション機能などについても考察することが当然必要である。周知のように、音声パターン理解の根底に横たわる思想は、人間の知覚の直接対象となる音声波形、あるいはそのシンボル化された記号系列からだけでは、メッセージの内容は理解できないという立場である。

例えば英語の [tu:] の発音記号に対し、正書法として単語 “to” “two” “too” が対応する。日本語の [い] の発音記号に対しては、位、委、井、伊、衣、亥、医、易、胃、威、異、律、意、緯と多くの漢字があるが、共に発音記号と単語とは一対多対応である。つまり使用単語、構文、意味の知識を利用しなければ表層構造の完全な文章に再現することはできないのである。

日本語に主語がないというのは、日本語ではコミュニケーションしている世界を既知と考えているからである。誰と誰が話の場にいるか、コミュニティの世界を頭に入れてから、言葉を発するのである。さらに日本語では、話し言葉の発音記号に対し、あまりにも多くの意味上の漢字が存在しているから、上例のような極めて単純な母音のオト「い」に対しても、言葉の一つとして捉え、続いた文脈の言語を記憶する左半球の

大脳に記憶をさせている。「詰」の音も、「轟」の声も、言葉とし、音楽として日本人が聞くのは、より長い文脈の中で、判断せざるを得ない日本語で育てられ、生活してきたからであろう。アルファベット系の人のように、Yes, No を文章の先に出して、頭脳の中で文脈の予想を容易にしている人々と、同じ言葉の長い系列の最後で、否定か肯定かが決まる日本語を話す人々とは、Context Sensitive な処理に大変な差がある。日本人が暗算が得意なのも当然である。英語の respectively, it, the former など、副詞、代名詞の理解のためには、やはり構文・意味の世界での対応づけを必要とする。このように、人間の言葉は、英語でも日本語でも、理解するためには、完全度は 50 歩 100 歩の表層構造である。相手の人の知識と関係させて初めて充分な解釈が成立することが判る。

コンピュータが自然言語の世界に深く入り込んでくればくる程、その入力、蓄積、検索、出力で自然言語のもっている深い構造を機械にも知ってもらう必要が生ずる。

私は昔から『漢字の機械への入力は、アルファベットに比較してハンディキャップがある。しかし、意味レベルの知識を活用して機械が理解するとき、連想記憶するときには、表意文字の漢字が表音文字より優れたことを示し、一挙に漢字かな混りの日本語が、コンピュータとの親和性で逆転することがあり得る』と話してきた。今や膨大な自然言語のデータベースと、文字列、音韻列で対話的にコミュニケーションする入力処理システムの時代に徐々に向ってきている。単にメッセージの表現が、ローマ字か、かな文字か、あるいは漢字であるかなどと言っているのは、商用の 3.5 世代から次の世代のコンピュータ時代のことであろう。研究としての次の世代のコンピュータでは、自然言語の入力、蓄積、処理のあらゆる面で、人間の脳で言えば右半球時代から左半球時代の構造をもったアーキテクチャの機械を設計することが必要だと思うのである。

(昭和 54 年 11 月 14 日)

† 本会副会長 京都大学工学部情報工学科