（『国語学』200集　2000・3・31）

# 日本語研究と電子化

辻井潤一

## 一　はじめに

計算機を使った言語研究が、ここ数年、急速に進展している。

計算機による言語研究は一九五〇年代の末から始まり、六〇年代の機械翻訳システムの隆盛期、六六年の機械翻訳に否定的なALPACレポートとそれに続く冬の時代、七〇年代の人工知能研究からの言語理解の研究といった具合に、いくつかの時期に分かれる。八〇年代は、筆者も関係した機械翻訳の国家プロジェクトなど、日本だけでなく、ヨーロッパを含めて、膨大な研究費が言語の計算機処理に注がれた[1][2]。

九〇年代は八〇年代の反動で、計算機による言語処理の研究が冷え込んだ時期であった。ただ、九〇年代も、後半に入ると、インターネットによる情報検索など、ふたたび計算機と言語の相互関係が緊密になり、多くの企業、大学での研究が急速に活発化している。

本稿では、言語の研究、あるいは、言語と情報技術の結びつきを、その言語が使われている社会との関係で考え、国語と情報技術についての私見を述べる。

## 二　社会の中の言語と計算機による言語研究

米国では、TIDES(Trans-lingual Information Detection, Extraction and Summarization)のプロジェクトが承認され、二〇〇〇年から開始される[3]。このプロジェクトでは、国家の安全に関連するテキスト情報をいち早く捉えて、その抄録を自動的に作る、といった実用システムの開発が目的とされている。米国は、日本語、主要なヨーロッパ諸語のテキストよりも、これまで計算機処理の観点からは比較的マイナーな言語とされてきた言語（たとえば、インド亜大陸の言語、アラブ・アフリカ世界の言語など）のテキスト処理に関心があるようである。明らかに、世界で唯一の超大国は、世界に覇権を唱えるための技術を開発することを目指している。

ヨーロッパでも、やはり言語処理技術に大きな力を注がれている。ヨーロッパ連合では、国境での出入国管理、輸入輸出の関税障壁を廃止し、通貨を統一することで、人・もの・金の連合内での自由な流通を達成した。次は、情報の自由な流通であり、その障壁である言語問題の解決が焦眉となっている。

障壁を除くのに、通貨の場合のように統一言語を作ることは問題

外である。文化の多様性と伝統の保持には、個別言語の存在が不可欠であることを、ヨーロッパ人はよく理解している。言葉の障壁を取り除く、あるいは、それを低くする鍵は、情報技術の活用にある。ここでは、従来の機械翻訳だけでなく、多言語での情報検索、分野の多言語の辞書作成といった、言語処理のためのインフラ作りに力が注がれている。インドでも、国内に多く存在する言語間の翻訳システムを作ることが、インドの将来の発展に不可欠との認識が強い。結局、言語の問題をどう捉えて、それを解決していくか、あるいは、その解決に計算機をどのように活用するかは、その社会がどのような方向に進む意志をもっているかに大きく依存する。

いま、日本では、社会として言語をどのように捉え、国語をどのような方向に導いていくかの意志に一種の混乱がある。そのことが、情報技術と日本語とをどのように結び付けていくかに明確な政策がない原因になっている。

## 三　国語の計算機処理

国語問題に系統的に取り組む国に、英国がある。一九九一年から九四年に作られた英語のコーパス、BNC(British National Corpora)は、一億語を超える巨大コーパスであり、オックスフォード大学出版局、Longman社などの出版社やランカスター大学などが、日本でいう科学研究費の補助を受けて作りあげたものである。この言語資源は、今後、単に英語学研究だけでなく、これをもとにした辞書など、英語教育の資産として活用されていくことであろう。実際、バーミンガム大学のコーパスから作られたコリンズ社の英語辞書Cobuildは、実際の言語使用の例を縦横に駆使した、優れた辞書になっている。また、そのCD版には、五〇〇万語からの実例コーパスがついている[4][5][6]。

このような優れた辞書は、英語教育の有効な資材として、英語の世界制覇に寄与することになる。このBNCに対して、米国の研究者は、やはり英語と米語の違いからBNCでは研究できないとして、ANC(American National Corpora)の作成を提案している。

このような国家的な言語政策は、もちろん、インターネットによる情報流通の国際化と切り離して考えることはできない。たとえば、「ポルトガル語のような大きな言語は、当然マイクロソフトの言語政策でも重要な位置をしめる」とビル・ゲーツの言葉に安心していたポルトガル政府が、マイクロソフトがブラジルのポルトガル語を念頭においていると知って、急遽、計算機による国語研究、あるいは、国語処理に研究費を出すことにしたなど、情報化社会の中での国語政策の重要性が急速に認識され、そのための施策が各国で取り上げられている。

直面する国際化の中で、国語をどのように守り、もり立てるかが、文化政策的な面だけでなく、経済面からも急速に重要になってきている。情報化社会は、情報を発信し、みずからのアイデアで主導権をとっていく社会である。いままでは、社会の持つ経済力が、その社会の言語の力を決定していた。いま、それが逆転し、言語の持つ力が、その言語を使う社会の経済的な力を決めていく。

## 四　提案と今後

このような状況の中で、日本社会には、日本語をどのようにプロモートしていくか、あるいは、多言語社会の中での日本語をどのよ

うに位置づけるかの議論が、まだ、十分に行われているようには見えない。問題が、英語教育、日本語コード体系の問題などに矮小化されているように見える。

国際化は、かならずしも、英語問題に矮小化されるべきものではない。インターネットの中に占める英語の地位の低下が、米国がTIDESを始める動機になっている。英語以外の言語と日本語の相互関係なども、もっと積極的に考えるべきであろう。教育をすることで、外国語に堪能な専門家を作るという発想だけでは乗り越えられない、未曾有の国際化が始まっている。情報技術の積極的な使用は、不可欠である。

また、日本語を力のある言語にするには、コード問題以上に、計算機の力を使った日本語研究、日本語教育のための言語資源の充実を計る必要がある。(7) これらは、いずれも研究者の個人的な努力ではなく、政策的な取り組みが必要な分野となっている。

国語の乱れも、実は、国内問題ではない。言語がコミュニケーションの手段である以上、それを可能にする系統性、保守性が必要であり、国内だけでなく国外でも日本語が使われていくためには、その現在態を客観的に捉え、標準化していく努力、言語としての表現力を意識的に保持する努力が必須となる。このような目的をもって作られた組織に、国立国語研究所がある。いま、その先見性をもって作られた研究所が、その真価を問われているように思える。

**[参考文献]**

(1) 和田 弘：機械翻訳、情報処理3-6、一九六二
(2) 辻井潤一：機械翻訳に何ができるか、「翻訳の方法」(川本、井上編)、東京大学出版会、一九九七

(3) http://www.darpa.mil/darpatech99/presentations.htm：米国防総省高等研究局のホームページで、TIDESの概要がある。
(4) http://info.ox.ac.uk/bnc/news/：BNCのホームページ
(5) http://www.athel.com/cobuild/cocd.html：CoBuild辞書のページ
(6) http://www.icp.grenet.fr/ELRA/home.html：ヨーロッパでの言語資源活動
(7) http://www.jeida.or.jp/gsk/gsk.htm：日本での言語資源活動

——東京大学大学院理学系研究科情報科学専攻、教授——