4K-1

計算機用自然言語辞書実用化の諸問題

村田賢一　須田直英　橋本三奈子
(情報処理振興事業協会)

1．はじめに

われわれは、さきに、計算機用日本語基本動詞辞書ＩＰＡＬ（ＩＰＡ Lexicon of the Japanese Language) (Basic Verbs)を試作し、その公開を行った。この辞書は、基本的な和語動詞約９００語について、まず統語・意味的規準に基づき細かく下位区分を行い、次にその夫々の下位区分毎に述語素、ヴォイス、ムード、テンス・アスペクト等を含む文法的情報を詳細に記述している。われわれは、この次のステップとして、上記下位区分の語義を述語論理を用いてフォーマルに記述する意味論的研究に着手している。

このような計算機用自然言語辞書をつくる試みは今後一層盛んになるものと予想されるが、実際にこの作業に従事した経験から言えば、これは一般に想像される以上に困難なものであった。

その理由は多数存在するが、その一つとして、計算機用自然言語辞書のための辞書学（lexicography）が甚だ未発達である、という点があげられよう。ＩＰＡＬ開発の経験については、実際的な問題を中心にさきに報告を行っているが、今回はこのような辞書学の形成にむけて、われわれの考察を述べる。

2．辞書の枠組みについて

2-1 辞書と文法

一つの言語の記述は一般的文法規則と語彙項目別の個別規則との総体として与えられるので、当然両者の間にはトレードオフの余地が大いに存在する。実用上の目的からいえば、後者即ち辞書により多くを分担させたやり方が良い、というのが一種の経験法則になっている。

いずれにしろ、辞書の枠組みを考える上で、何らかの文法体系を想定しないわけにはゆかない。実際上は文法体系の選択がなかなか難しい問題なのであるが、ここでは適切な文法体系が与えられたと仮定する。

このような文法の中の規則の一般形として、「語彙項目Ａを含む句又は文の～に関する規則は～に関する当該語彙項目の分類番号がｉで与えられると、f(i,A)で与えられる。」の如きものが想定される。

例えば、ある動詞Ａを含む直接受動文を作る規則は、ＩＰＡＬにあるような、能動／相互／中動／受動という分類情報及び格形式の交替パターンに関する情報をもとに、容易に与えることができる。

また、テイル形式の意味については、ＩＰＡＬでは夫々の動詞のアスペクチュアルな分類を行っているが、動詞だけではアスペクトの多義性の解決が出来ないので、アスペクト判定規則は、関与する名詞句や副詞句の素性等を考慮に入れた複雑なものになる。いずれにしろ、動詞の分類番号が重要な役割を果す。

このように考えると、「辞書における記述は本質的に分類作業である」ということができる。

2-2 分類方法が満たすべき諸要因

このような辞書に於る分類が文法規則の記述に有効なものであるべきことは当然の理論的要請であるが、他方、辞書を実際に作る立場からは、充分実行可能な分類法であることが望まれる。というのは、言語学理論の方から要請される分類規準は時として概念的規定になっており、作業者が夫々の理解した所に従って分類作業を行うと辞書の品質低下の原因になる。従って、テストを用いるなど、より操作的な、判定規準の明確な分類法がよいのは当然だが、実際上はいろいろ細かな点まで規定しないと、良質の結果を安定して得ることが困難である。

テストの付帯条件の例をあげると

- 動詞とその格要素の名詞句だけで判定する場合、名詞句の素性上の制約（個体、種、物質、単数／複数など）
- 副詞句との共起をテストする場合で、当該副詞句に意味上の下位区分を必要とする場合は、そのどの下位区分によるテストかを操作的規準で示す。
- 文脈を考えに入れてのテストを行う場合は、どのような文脈を使ってよいのかということ。（例ではメタフォアは排除するという条件をつけた場合を考えると、その境界線の示し方が問題になる。）

3．記述者の資質と記述方法の問題

一つの言語を記述する上での困難な点は、個々のインスタンスは直接触れることができるが、一つの言語それ自体には触れようがないことである。ましてや、個々の記述者を採用して記述させる場合、記述者個人の諸々の言語的背景に基づく干渉現象を如何に排除するかが重要な課題になる。

このような干渉現象の要因を若干列挙すれば：（イ）方言的要素，（ロ）時代的・年代的要素，（ハ）生活環境，（ニ）知識・教養，（ホ）外国語，（ヘ）流行現象，等がある。誤解がないように若干説明をつけ加えると、（イ）は、長年日本語の研究に従事しているような人であっても、本人の生育歴上最も影響の強かった方言の干渉を避けることが困難であるということである。意外と問題なのは（ホ）であって、例えば、受身文の判定を行わせると国文科の学生と英文科の学生では結果が異なるというようなことは頻々経験することである。これはまた意味を扱うとき、日本語の中の漢字は古代中国語からの借用物であるから、頻々etymological fallacyの原因となる。

さて、このような諸々の干渉現象の対策として、第1図のようなアンケート用紙を作ってみた。これは、問題の表現について客観的なデータを得る上で有用であるのみならず、記述者に自己の判断結果に対する反省材料を与えることによる教育的効果も持っている。

第1図
アンケート用紙

表　現	
意　味	

	上記の表現については				
インフォーマント	その意味で理解していた		その意味では理解していなかった		
	自分でも使っている	自分は使わないが人が使っているのは見聞した	人が使っているのを見聞したことがあるような気がする	聞いたことも見たこともない	別の意味で理解している（具体的に）
1					
2					

4．あとがき

最近は電子化辞書という言葉に象徴されるように、計算機用自然言語辞書への関心が大いに高まっている。しかし、実際上の作業はなかなか困難であって、そのような経験を踏えた辞書学の確立が望まれる。われわれはＩＰＡＬ試作の経験から得られた結果を若干ここに報告したが、まだまだ解らないことも多い。例えばコーパスの利用であるが、一体どの位のサイズのコーパスがあればよいのか、ということも明らかになっていないし、理論上要求されるであろうサイズのコーパスを如何に有効に作り、利用するかにも多くの問題がある。

5．謝辞

学際研究の性格上、実に多くの方々の御協力を得た。特に、御指導いただいたコンサルティング委員会の先生方、共同研究に参加して下さったワーキング委員会及び臨時ワーキング委員会の諸氏およびこの研究の機会を与えられた通産省情報処理振興課ならびに情報処理振興事業協会の各位に深甚の謝意を表します。

参考文献

(1)村田、村木、須田、橋本：計算機用日本語基本動詞辞書について　―その分析と評価―，
情報処理学会第３２回（昭和６１年前期）全国大会講演論文集pp.1577-1578.
(2)計算機用日本語基本動詞辞書説明書，情報処理振興事業協会，昭和６１年３月.