

知識情報処理特集 ————— ②

知識情報処理の展望

- 2-1 自然言語処理の立場から
- 2-2 パターン理解の立場から
- 2-3 神経科学の立場から
- 2-4 知識情報処理と認知工学
- 2-5 通信工学の立場から

2—1 自然言語処理の立場から

田中穂積

田中穂積：正員 東京工業大学工学部情報工学科

1. はじめに

自然言語処理の一つの目的は、自然言語を理解して応答するシステム（自然言語理解システム：Natural Language Understanding System）を作成することにある。人間に近い言語理解能力をもつシステムの実現に一步でも接近しようとする努力が、これまで人工知能の研究者によって粘り強く続けられてきた。我が国の第五世代コンピュータ計画でも、自然言語理解システムは、知的インタフェースシステムの重要な研究テーマの一つとして取り上げられている。これは、人間とコンピュータとの間の使用言語の相違に基づくセマンティックギャップを解消し、人間とコンピュータとの間に新しいインタフェースを確立しようとするものである。

本特集号の知識情報処理という術語は、第五世代コンピュータシステムを機能面から特徴付けるために使われたのが最初であったと記憶している。筆者は、第五世代コンピュータの知的インタフェースシステム計画立案段階で、「知的インタフェースシステムは、知識情報処理システムの一構成要素ではあるが、それ自身、典型的な知識情報処理システムである」と述べたことがある [Tanaka 82]。

2. 自然言語処理を巡る諸知識

自然言語処理を行うための知識を大きく分けると次

の二つがある。

[1] 言語学的な知識（辞書に記載された知識や文法）

[2] それ以外の知識（常識等）。

[1] については言語学者がさまざまな角度から研究している。とりわけ文法については、言語学者が詳しく研究しており、言語学的な理論として多くの知見が得られている。これらは主として文法規則に関する知識であるが、現在それらを部分的に利用した自然言語処理システムも作成されている。

文法規則と対になる知識として重要なものに、語彙（い）のもつ知識がある。語彙とは単語の集りのことであるが、これまで語彙の研究は、その記述量が膨大であること、地道な用例分析の積み上げを長年にわたって多数行う必要があることなどから、言語学者の多くはこれを泥沼とみなし敬遠してきたように思われる。

ところが最近事情が少しずつ変りつつある。言語学の新しい潮流として非変形文法の枠組みが注目されている⁽²⁾。そこでは、各語彙項目にかなりの情報を持たせておくことにより、これまで説明が困難であったいくつかの言語現象をきれいに説明し直すことが可能になることが明らかになってきたからである。この新しい言語学では、語彙に書かれる知識が、言語学的な現象を説明する鍵（かぎ）になる、という考え方に基づいている。従来の言語学では、語彙に書かれる知識の量が貧困であったために、理論的な困難さが生まれてき

たとする考え方である。とはいえ、各語彙項目にどのような知識をどのような形式で記述すべきかについての研究が言語学者によって十分になされているとはまだいえない。

一方、自然言語処理の研究者は、自然言語理解システムや機械翻訳システムの開発を通じて、辞書にどのような情報を盛り込むべきかに依存して、システムの性能が大幅に変ることに気付いていた。このような辞書は、我々が日常使う辞典とは異なり、計算機に可読な形式をしていなければならない。最近我が国で電子化辞書プロジェクトが始り、大規模な電子化辞書の構築に向けての研究開発が行われている。知識情報処理を背後から支えるプロジェクトとして今後の成果が期待される。これについては、3.3 で述べる。

[2] の常識は、高度な自然言語処理を行う場合に必要になる。このことは、例えば言外の意味を理解したり、話者の意図を理解⁽⁴⁾する場面を考えればわかるだろう。しかし問題は、常識がどのようなものであるかについて、哲学者、心理学者、言語学者等によって断片的な研究はなされているものの、体系的な研究がなされているとはいえないことである。

言語学者の多くは、この常識と言語に関連した問題は語用論 (pragmatics) の問題であり、現在の文法理論の中心的な研究課題ではないとしてそれ以上深く追求しない。哲学者も、自然言語処理の立場から利用可能な常識という観点から常識の研究を進めてきたわけではない。最近になって、状況意味論⁽⁴⁾などで、自然言語処理の研究にも有効と思われる理論が生まれつつあるが、今後のより一層の研究が必要である。

自然言語処理システムを作成する立場からは、常識の問題が困難であるからといって避けるわけにはいかない。しかし現在の研究レベルでは、この問題を一般的に解くことができる段階にはない。そこで、対象分野を極めて狭くとり、限られた範囲の自然言語処理を

行うシステムを考えざるをえない、というのが現状である。

3. では、自然言語処理システムを作成する上での諸知識の表現と推論の問題を論じてみたい。そして 4. では知識情報処理システムとしての自然言語処理について述べる。

3. 知識表現と推論

ここで自然言語処理システムを作成する上での望ましい知識表現とはどのようなものかを考えてみたい。筆者は、知識と推論とは背中合せでなければならないと考えている。推論機構の裏付けのない知識は不完全な知識であるといって良い。この考え方は近似的に図 1 の等式で表される。本章では、文法規則と常識の表

$$\text{知識表現} = \text{表現形式} + \text{推論機構}$$

図 1 知識表現の等式

望ましい知識表現は推論機構の裏付けがなければならない。

現方法を、図 1 の推論機構を行うプログラムを作成する必要があるかどうかという観点から比較する。最後に辞書の問題について述べる。

まずはじめに、自然言語処理を行う場合に必要になる知識の一つである文法規則と推論の問題を論じる。次に、常識の表現と推論の問題を論じる。そして、ロジックプログラミングの枠組みで知識表現を行うことの利点を明らかにする。

3.1 文法規則と推論

自然言語処理で最も良く使われる文法規則記述形式は、文脈自由文法規則に補強を施したものであろう。これは例えば次の形をしている。

$$S \rightarrow NP VP \{ \text{補強項} \}.$$

文 (S) は主部 (NP) と述部 (VP) とからなる、という知識を表したものである。

文脈自由文法規則を用いて、与えられた文を構文解析するプログラムは、パーザとよばれている (図 2)。パーザにはさまざまなアルゴリズムが考案されているが、パーザは言語的な知識 (辞書と文法) に対するインタプリタとして働く。いい換えると、言語的な知識を用いて構文解析結果を推論するプログラム (推論機構) であるとみなすことができる。



田中 穂積
(正員)

Natural Language Processing in Knowledge Information Processing. By Hozumi TANAKA, Member (Faculty of Engineering, Tokyo Institute of Technology, Tokyo).

昭 39 東工大・理工・制御卒。昭 41 同大学院修士課程修了。同年電気試験所 (現電総研) 入所。昭 58 東工大・工・情報・助教授。現在、教授。OS, 人工知能, 自然言語処理の研究に従事。工博。

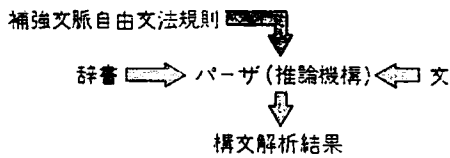


図2 パーザと構文解析

文法規則と辞書とを用いて、与えられた文を構文解析するプログラムをパーザという。

図2に示すように、これまでの自然言語処理システムでは、構文解析を行うためのパーザ(推論機構)の作成が大きな課題であった。

ところが Colmeraure の開発した Metamorphosis Grammar⁽⁴⁾ を発展させた Pereira⁽⁵⁾ の DCG (Definite Clause Grammar) の考え方では、拡張文脈自由文法の一形式である DCG で書かれた文法規則と辞書を、それと一対一に対応する Prolog プログラム (Horn 節) に変換する。変換後の Prolog プログラムに自然言語の文を与えて実行すると、トップダウンで縦型探索 (depth first) による構文解析を行うことができる。

この方法によれば、従来、自然言語処理で必要としたパーザの作成が不要になる。Prolog のインタプリタに組み込みの機能をそのまま用いて、パーザを代用することが可能になるからである。

しかしこの方式には一つの大きな問題がある。それは、左再帰的な文法規則があると、構文処理の過程で無限ループに陥るという問題である。この問題はボトムアップな構文解析を行うことにより避けることができる。

電子技術総合研究所の松本らの開発した方法によると、DCG の形式で書かれた文法規則を、ほぼそれと一対一に対応する Prolog プログラムに変換し、この Prolog プログラムに自然言語の文を与えると(構文解析と意味解析とを融合した)ボトムアップで縦型探索の構文解析を行うことができる⁽⁶⁾。これは、BUP システムとよばれている。

Prolog を用いた構文解析の基本的な考え方を図3に示す。

ここで次の注意をしておく。図2の方法では、意味解析と構文解析とを同時に進めるためには、特別なプログラムをパーザに組み込まなければならなかった。ところがパーザのつくりによっては、この特別なプログラムの組み込みが必ずしもうまくいかないことが多かった。そのため図2の方法では、意味解析と構文解析

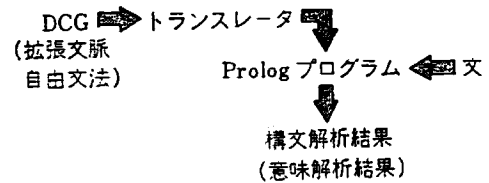


図3 Prolog を用いた構文解析

Define Clause Grammar (DCG) の形式で書かれた文法規則を、ほぼそれと一対一に対応する Prolog プログラムに変換し、この Prolog プログラムに自然言語の文を与えると構文解析ができる。

とを融合化したパーズングが課題になっていた。しかし図3の方法では、DCG の拡張部(補強部)に Prolog プログラムを埋め込み実行するだけで、簡単に意味解析と構文解析とを融合することができる。これは心理学的にも妥当な解析モデルであるといえるが、詳細は例えば文献(8)を参照してほしい。

Prolog はロジックプログラミング言語の典型であるから、図3に示したことはロジックプログラミングの枠組みと自然言語処理との整合性の良さを示唆している。

3.2 常識と推論

常識をどのように表現するかは難しい問題であるということを、2. で述べた。ここでは比較的単純な常識を表現することを考える。

知識表現で良く使われる形式に、セマンティックネットワークとフレーム⁽⁷⁾がある。簡単な常識を、セマンティックネットワークで記述したものを図4に示す。セマンティックネットワークで記述した知識を、図1の立場から見るとどうなるだろうか。この種の知識表現では、それを解釈するためのインタプリタを別途必ず作成しなければならないことに注意しよう⁽⁸⁾。推論はこのインタプリタを介して行われる。

従って、フレームやセマンティックネットワークで表現された知識は、使用者個々が適当な推論機構をインタプリタに組み込まない限り、推論機構の裏付けを持たない不完全な知識であるということになる。このような知識の問題点は、それをどのようなものとみなすかは、使用者個々の作成するインタプリタのつくりに依存することである。そのため表現されたものの形式は同じでも、その意味が使用者に依存することになり、外部からみて不明確になる。

以上を図示すれば図5のようになる。

(b) の部分は、セマンティックネットワーク、フレ

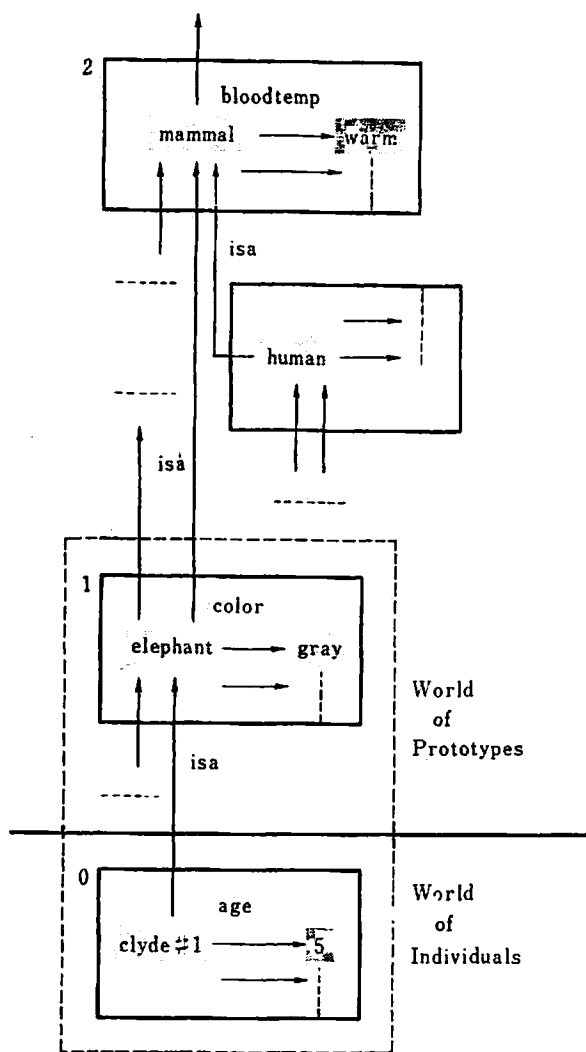


図 4 セマンティックネットワークによる知識表現例

意味の世界を概念間の関係としてとらえる考え方を、セマンティックネットワークとよぶ。

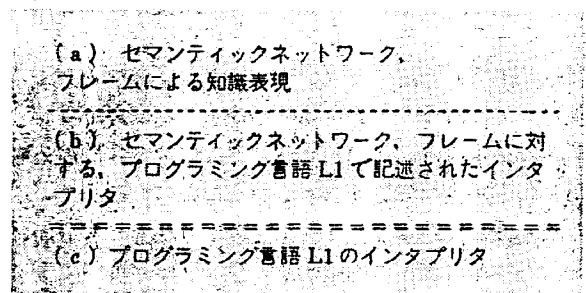


図 5 セマンティックネットワーク、フレームによる知識表現

セマンティックネットワークやフレームによる知識表現では、それを解釈するプログラムを作成して推論を行わなければならない。

ームによる知識表現を使うために個々のユーザが作成するインタプリタで、図 1 の推論機構に相当する部分である。一般に (b) を作成する労力は無視できない。(c) はプログラミング言語 L1 に対するインタプリタであるから、システムに組み込まれている。

先に述べた知識表現に対する等式からすれば、形式的な推論機構の裏付けのある述語論理による知識表現は、望ましい知識表現の一つであるといえる。特にロジックプログラミングの枠内で知識表現の問題を考えることは意味がある。なぜなら、一定の形式の述語論理で表現された知識に関する推論は、すべてロジックプログラミングシステムに組み込みの機能を利用して行うことが可能になるからである⁽¹¹⁾。

その様子を図 6 に示す。図 6 (b) が「なし」と記述してあるのは、例えばロジックプログラミング言語の典型である Prolog に組み込みの推論機構が利用できるからである。

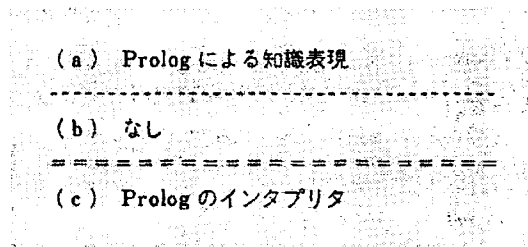


図 6 Prolog による知識表現

Prolog により知識表現を行えば、Prolog に組み込みの機構を利用した推論を行うことができる。

3.3 辞書

自然言語処理を行うためには、辞書が必要になる。辞書には、文法的な情報のほかに 3.2 で述べた常識などが記述されることになる。後者は高度な自然言語処理システムを開発する場合に必要な。次の世代の自然言語処理システムを研究開発する目的で、これまでの研究成果を集大成し、新しい考えを組み入れた大規模な電子化辞書開発プロジェクトの構想が具体化している。これについて説明する。

一般に人間に可読な辞書であっても、システムに可読で使用可能な形式であるとはいえない。システムに可読で使用可能な大規模な辞書の開発には膨大な時間と労力が必要なのが推察される。本年通産省は電子化辞書プロジェクトを発足させた。

電子化辞書プロジェクトでは、日本語と英語を対象とし、まずマスタ辞書を開発する。マスタ辞書には基

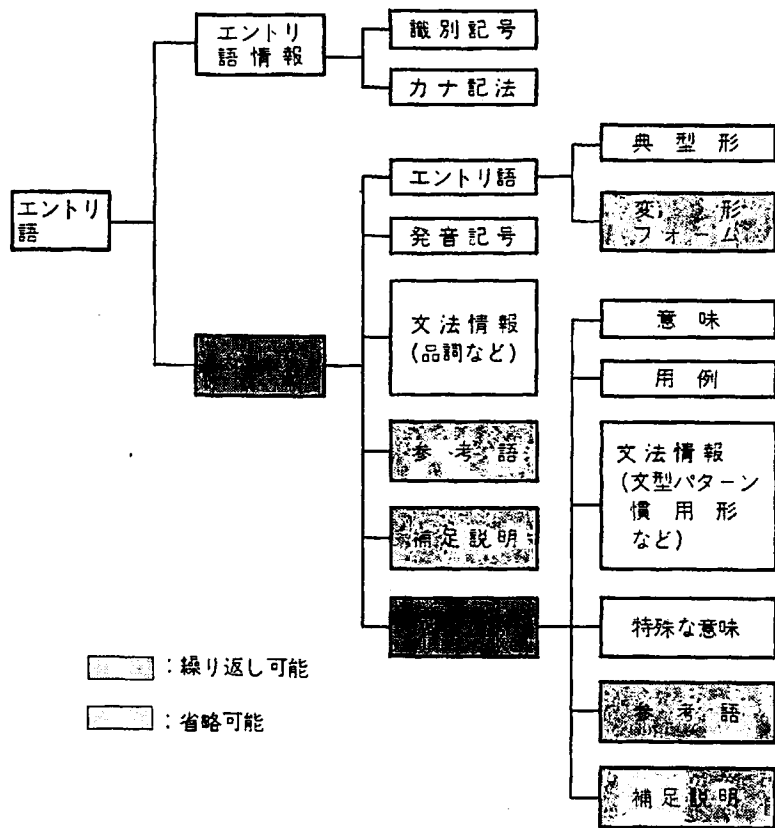


図7 日本語辞書の形式

マスタ辞書の各項目はモジュール化された記述の階層構造になっている。

本語辞書と専門用語辞書とを含む。概念辞書はマスタ辞書の語の意味をコンピュータに理解可能な形式で表現したものであり、概念と概念の関係を表すソーラスは、ここに含まれる。

マスタ辞書の形式については第五世代コンピュータプロジェクトで検討してきたものをベースにしようとしている。これを図7に示す⁽¹²⁾。

電子化辞書プロジェクトは9年計画であるが、電子化辞書の作成支援ツールとして、また新しい専門用語の管理登録を行う運営支援用ツールも作成されることになっている。このプロジェクトは地味ではあるが、高度な自然言語処理システムを作成するための基本データとして将来大きな役割を果たすものと期待される。知識情報処理を背後から支える重要なプロジェクトである。

4. 自然言語処理システムと知識情報処理システム

知識情報処理システムとしての自然言語処理システム

ムの構成を図8に示す。

知識情報処理システムの立場から自然言語理解システムを眺めると、中央に巨大な知識ベースがあり、そこに記憶されている知識を利用しながら言語理解を行い応答するシステムが自然言語理解システムであるということになる。その意味からすると、図8の知識ベースの枠はもっと大きく書いても良かったかもしれない。

図8からもわかるように、言語理解システムには次の三つの異なる側面がある。

- [1] 文の解析
 - [1.1] 構文解析 (形態素解析を含む)
 - [1.2] 意味解析
 - [1.3] 文脈解析
- [2] 推論 (問題解決)
- [3] 応答文生成

このうち [1] と [2] は言語理解に直接関係する側面である。[3] は我々が言語を話したり書いたりする場面で必要になるものである。

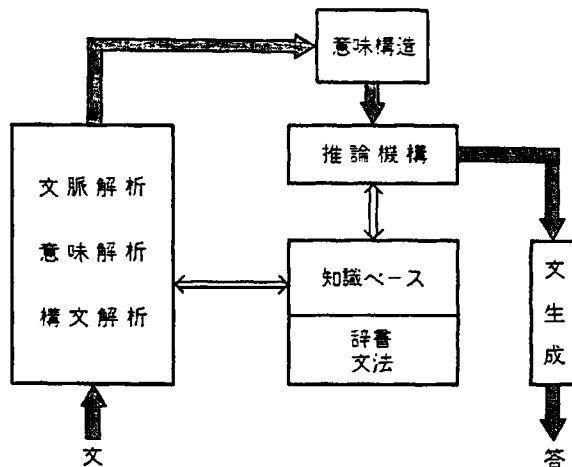


図8 知識情報処理システムとしての自然言語理解システム

知識情報処理システムの立場から、自然言語理解システムを眺めると、中央に巨大な知識ベースがあり、それを利用しながら言語理解を行い応答するシステムということになる。

限られた領域での自然言語による応答を行う自然言語理解システムでは、[3] はこれまでさして重要視されていなかった。[1] と [2] が十分でない、とても自然言語を理解するシステムの作成はおぼつかない、ということは明らかだろう。そのため、これまで [1] と [2] に研究のウェイトが置かれてきたと考えられる。特に [1] では解析結果にあいまいさが含まれ、一意に決まらないことが多いのが問題である。

[3] の問題は自然言語理解システムの研究ではなく、最近機械翻訳システムの研究でクローズアップされてきている。機械翻訳システムの性能は、[3] の能力に大きく依存するからである。機械翻訳システムの研究を契機に、今後の研究の発展が期待される。

知識情報処理システムの中核をなす知識ベースに再び話題を移すと、これには次の三つの問題がある。

[A] 知識表現

[B] 知識の効率的な検索と更新

[C] 知識獲得

[A] の自然言語処理に関連した知識表現の問題は人工知能の研究の重要な研究課題である。言語学的な知識をどのような形式で表現しておくかは、言語学者による研究成果がある。例えば多くの言語理解システムでは、言語学という文脈自由文法規則 (context-free rule) に補強を施したものが利用されている。辞書項目に記述される知識も、品詞や言語学的な素性だけでなく、意味解析を行うための知識が利用されている。以上については 3. で簡単に述べた。

限られた範囲の自然言語を対象とした自然言語理解システムを研究している間は、扱う辞書項目の数が少なく、辞書項目検索に要する時間は問題にならない。しかし機械翻訳システムのように、数万の辞書項目を対象とする自然言語処理システムを作成する場合には、辞書項目のすべてを一次記憶に置くわけにはいかず、そのほとんどは二次記憶に置かなければならない。そのため辞書項目の検索時間が問題になる。翻訳のための基盤知識である辞書項目を、どのように構成しておくかは重要な問題である。また絶えず新しい用語が作り出される科学技術文献を対象にした自然言語処理システムを作成する場合には、辞書項目の効率的な更新が可能でなければならない。これが [B] に述べた問題であり、これに関連して、最近の電子化辞書プロジェクトの考え方を 3.3 で説明した。

筆者は、[B] の問題の解決には、知識の効率的な検索・更新を専用に行うマシン (ハードウェア) を開発

する必要があると考えている。知識ベースマシンの開発である。知識ベースマシンのうち、最も単純で良く研究されているものは関係データベースマシンであろう。関係データベースマシンを開発することは、自然言語処理との関連で、膨大な辞書項目を扱う辞書マシンとして役立つものと期待される。関係データベースマシンの具体的な応用例として辞書の問題を考えることは意味があると思う。

[C] の知識獲得の問題も、知識ベースが小規模のうちにはさして大きな問題とはならない。しかし知識ベースとして大規模なもの (例えば辞書) を開発しなければならない場合には問題である。自然言語処理システムが使用可能な辞書の形式を定め ([A]), その形式の辞書を開発しなければならないからである。

この問題を解決するための一つの方法は、我々が使用する辞書 (国語辞典、英和辞典、和英辞典など) から、システムに可読な形式の知識を自動的に抽出獲得する方法を開発することである。このような研究は知識獲得の研究に含まれる。しかし知識獲得の研究は緒についた段階にある。学習の問題とも関連する困難な研究課題であることが知られている。一層の研究が必要である。

5. おわりに

我々は頭脳のなかの限られた資源を用いてさまざまな自然言語処理を行っている。問題ごとに異なる自然言語処理のためのプログラムが動作すると考えることもできよう。しかしその場合であっても、それらすべてに共通する部分があって、それを中核としてさまざまな自然言語処理を行っていると考えられることは自然であろう。その中核の部分とロジックプログラミング言語の持つ基本計算機構とを重ね合わせてみることはできないだろうか。文献(13)による水平的な処理を行うための基本計算機構としてである。4. までに述べてきたように、ロジックプログラミング言語の基本計算機構をそのまま用いて、従来個別に行われていた構文解析、意味解析などの問題を統一的に扱うことができる。ロジックプログラミングの枠組みでこの多様な問題を統一的に扱えるという事実は筆者の予想を越えるものがあつた。

以上のことから、文献(13)の水平的な処理を行うための機構とロジックプログラミング言語の基本計算機構とを重ね合わせてみるという問題に、哲学的、心理学的な検討を加える価値が全くないとはいえないと思う。

たとえ結論が否定的であったとしても、人間の持っているものと、どの点が異なるかを明らかにすることができれば、それだけでも意味があると思うのであるが。

文 献

- (1) H. Tanaka, et al. : "Intelligent Man-Machine Interface", in Ref. (14), pp. 147-157 (1982).
- (2) P. Sells : "Lectures on Contemporary Syntactic Theories", CSLI, Stanford University (1986).
- (3) A. Joshi, B. Webber and I. Sag : "Elements of Discourse Understanding", Cambridge Univ. Press (1981).
- (4) J. Barwise and J. Perry : "Situations and Attitudes", The MIT Press (1983).
- (5) A. Colmeraure : "Metamorphosis Grammar, in Bolc (ed) : Natural Language Communication with Computers", Springer-Verlag, pp. 133-190 (1978).
- (6) F. Pereira, et al. : "Definite Clause Grammar for Language Analysis—A Survey of the Formalism and a Comparison with Augmented Transition Networks", Artif. Intell., 13, pp. 231-278 (1980).
- (7) Y. Matsumoto, et al. : "BUP—A Bottom-up Parser Embedded in Prolog", New Generation Comput., 1, 2, pp. 145-158 (1983).
- (8) 田中徳積, ほか : 自然言語処理における Prolog", 情報処理, 25, 12, pp. 1396-1403 (昭 59-12).
- (9) D.G. Bobrow, et al. : "An Overview of KRL-O", Cognitive Science, 1, 1, pp. 3-46 (1977).
- (10) N.J. Nilsson : "Principles of Artificial Intelligence", Tioga (1980).
- (11) J.W. Lloyd : "Foundations of Logic Programming", Springer-Verlag (1984).
- (12) T. Ishiwata, et al. : "Basic Specifications of the Machine-Readable Dictionary", TR-100, ICOT (1985).
- (13) J.A. Fodor : "Modularity of Mind", The MIT Press (1983).
- (14) T. Moto-oka : "Fifth Generation Computer Systems", North-Holland (1982).