

特集/計算言語

## 計算言語学のすすめ



田中 穂積

### 0. はじめに

計算言語学 (computational linguistics) はコンピュータを用いて自然言語を研究すること及びその応用に関する研究を行う学問である [長尾 83] [Grishman 86]. 相互に関連し合っているが、プログラム言語 (計算機言語) の研究を行う学問ではない。計算言語学はその応用である機械翻訳システムが実際使われ始めてきていること、コンピュータと人間との間で円滑な意志疎通を行うための言語理解システムの研究 [Winograd 76] の重要性が理解されてきたことから、コンピュータの将来にとっても重要な学問であると認識されてきている [田中 88]. これらは応用計算言語学ともいう分野である。筆者の所属する工学部にも応用計算言語学に興味を示す学生が多数いる。

計算言語学には自然言語とその応用に関する研究の他に、人間の知のメカニズムを研究する学問である言語学、心理学、哲学と関連する研究がある [Gardner 85]. 以下ではこれらの研究者に対して計算言語学のすすめを述べてみたい。計算言語学の目指す問題が言語学、心理学、哲学の問題

とどのように関連しているかを明らかにし、応用計算言語学ではない計算言語学とは何かを説明してみたい。

### 1. 言葉とコンピュータ

コンピュータの誕生は人間社会に計り知れない影響を与えてきた。長い歴史の中で人間が初めて手にいれた、自立する精神を持つかのごとき印象を与える(?)マシン、それがコンピュータである。コンピュータが計算機とも呼ばれることから分かるように初期のコンピュータは、狭義の意味での計算、言い換えると四則演算を中心とする数値計算を高速に行うマシンであった。これは科学技術の分野に多大な影響を与えた。今やコンピュータは科学技術の発展にとって不可欠な存在になってきている。

確かにコンピュータの数値計算能力の素晴らしさには目を見張るものがある。しかしコンピュータの行う計算は数値計算に限られるわけではない。むしろ数値以外の記号を操作すること、言い換えると記号計算の能力を持つことがコンピュータの本質であるといつてよい。数値計算はどちら

かというは無味乾燥で人間臭さが感じられない、数値計算よりはるかに人間臭い計算にコンピュータは能力を発揮する可能性がある。

たとえば数値以外の記号として言葉がある。言葉は人間臭い記号の典型だろう。自然発生的に生まれた言葉は自然言語と呼ばれている。言葉は秩序ある記号の系列である。英語には英語の、そして日本語には日本語の秩序がある。言葉を理解し使用することは、秩序に従って記号を操作することである。こうした秩序のことを我々は文法と呼んでいる。秩序に従って述べた言葉を扱うことは、我々の高度な記号処理能力に依存している。

言葉に関する学問に言語学がある。言語学は言葉に内在している秩序を明らかにする学問である。ところで言語学の面白さと難しさは、言葉の秩序の解明が、我々の持つ高度な記号処理能力の解明と密接に関連している点にある。さる著名な言語学者が、心理学の一分野として言語学を位置づけているのはこのような事情によるのだろう。言語学の研究から明らかのように、言葉という記号の計算は、数値計算とは比べものにならないほど豊富な構造と潜在能力を持っている。コンピュータの本質が数値計算以外の記号操作にあり、言葉が秩序ある記号の系列であり、言葉を理解し使うことは記号操作の典型であることから明らかのように、言葉とコンピュータは生まれながらにして相性がよいと結論することができる。

コンピュータを用いて言葉の秩序を研究したり、コンピュータで言葉を解析したり作り出したりする研究が最近活発である。前者は言語学の研究にコンピュータを用いることであり、後者は前者の研究成果を応用することである。1960年代に生まれた計算言語学は、この両者を扱う学問である。2章では言語学と計算言語学との関連を説明する。言語学と計算言語学は研究の姿勢が若干異なる。これまでは計算言語学は言語学からの影響を受けてきた。しかし計算言語学の研究成果が言語学に貢献する機会が今後増えることが予想される。言語学でも計算言語学の影響を受けたと見られる理論が生まれるようになってきた。3章では

言語を理解する過程で生じる問題を巡って、心理学と計算言語学との関連について説明する。4章では言語の理解は、哲学的な問題とも関連することを説明する。

## 2. 計算言語学と言語学

前章では、コンピュータを道具として言葉を研究することができるということ、そしてそれは計算言語学の主要な一分野であるということの説明した。本章では言語学とコンピュータとの関連を次の二つに分けて論じる。コンピュータは、

- [1] 大量の言語データの蓄積と高速な検索、統計的な分析を行うことができる、
- [2] コンピュータの内部に言語理論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言語理論の妥当性を検証し、それにより言語理論を精密化したり修正することができる。プログラムによりコンピュータは検証しようとする言語理論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理論を作り出すことができる。

言語学の研究で言語資料をカード化したものをよく用いる。確かにカード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。そういう利点は確かにある。しかしカードの枚数と内容量が増えるにつれて検索に要す時間も増えるだろう。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現実には人手を大幅に軽減し高速化することができる。将来半導体集積技術の進歩により、カード化した資料が辞書程度の重さと大きさのコンピュータにより、電車の中でも自宅の居間でも、必要なときいつでも情報を引出し分析することができる時代が到来すると思われる。

このことは、コンピュータの記憶容量が最近飛躍的に増えていることから窺われる。価格も極めて安価になってきているだけでなく、半導体集積技術の進歩により、4 M ビット容量を持つ数ミリ四方のチップが販売されようとしている。ちなみに三省堂の新明解国語辞典は、およそ 64 M ビットの記憶容量があれば十分収容できる。現在でも研究段階では一チップ当り 64 M ビットから 256 M ビットの集積度が達成できるという。コンパクトディスク (CD) を用いると記憶容量は飛躍的に増える。これらの小型化と軽量化は時間の問題であろう。

言語学の研究には具体的な用例の分析が重要である。言語学者の中には用例分析よりむしろ内省と直観が重要であるとして、用例分析を軽視する傾向があると聞く。これは言語学者が言語分析用の資料の収集に追われて、それだけで研究生活の大半を過ごすことのないよう警鐘を鳴らすという意味があったのかも知れない。以前はたとえコンピュータを使用しようと思っても、コンピュータは一部の研究者が独占的に使う道具であり、コンピュータ自身の使い勝手が悪いという問題が確かにあった。

しかし最近これらの障害は取り除かれつつある。コンピュータは非常に安価になり、言語研究者がいつでも使えるようになってきた。コンピュータを使って大量の言語データを蓄積し、それらを分析することは以前とは比較にならないほど容易になっている。言語学者がそれを利用しない手はないと思う。それを利用しつつ内省と直観を働かせた言語研究を行うべきだろう。情報検索に関する計算言語学により、近い将来それが可能になるだろう。統計的な分析はコンピュータのもっとも得意とするところであるから、単なる頻度調査は研究にはなくなるだろう。用例の分析もそうだが、そこで求められるのは、大量の言語データを扱う行為そのものではなく、調査結果をどう解釈するかに対する研究者の洞察力である。

何年か前にハーバード大学の言語学科の学生が夏休み 3 週間ほど筆者の研究室に滞在したことが

ある。彼はサマースクールで計算言語学の授業を受け、いくつかのプログラム言語に熟知しており、簡単な機械翻訳システムを授業で作成したといていた。海外の全ての言語学科にこのようなカリキュラムがあるとはいえないが、わが国の言語学科でも、コンピュータに関する教育に真剣に取り組むべき時期にきている。

ここで計算言語学の研究と、これまでの純粋な言語学の研究との違いを簡単に説明する (表 1)。それにより [2] で述べたことが一層はつきりすると思われる。また [2] で述べたことが直ちに実現可能ではないということも分かるだろう。ただし以下に述べることは現在の時点での両者の研究姿勢の違いであって、それが将来ドラスティックに変わることは十分ありうる。

計算言語学	言語学
理論/応用	理論
厳密性重視 (アルゴリズム)	直観的な説明
言語理解過程	言語生成過程
統語論/意味論の連携	統語論の自立性重視
分析対象大量	分析対象少量
.....	例外事象に興味
曖昧さ解消	.....
	ユニフィケーション文法
	レキシコンの重視

表 1 計算言語学と言語学

表 1 について少し説明しておきたい。計算言語学では理論だけでなくその応用も考えている。1 章では後者を応用計算言語学と呼んでいた。自然言語を理解して応答するシステムや機械翻訳システムを開発することは、応用計算言語学の大きな目的である。このことは計算言語学国際会議を組織しているメンバーのほとんどが、過去において機械翻訳プロジェクトに携わった経験を持っていることから窺える。これらのシステムはプログラムとして実現されるため、計算言語学には言語処理を行うプログラムやアルゴリズムに関する理論も含まれる。言語学の理論には普通これらは含まれない。

言語学の理論は人間を対象にして構築するので、人間が理解できる言語理論であればよい。そのため言語理論は人間の直観に訴えた表現を用いて理論化を行うことがほとんどであるといつてよい。ところで人間の直観に訴える表現の最たるものは自然言語であるが、自然言語を用いた言語理論をコンピュータ上にモデル化しようとするとはしばしば問題を起す。

たとえば、「一般に、文の前半は古い情報を担い、文の後方は新しい情報を担う」ということがいわれている。これは言語によらない普遍的な原則であると主張されることがある。ところが、「文の前半/後半」とはどこからどこまでかとか、「古い情報/新しい情報」とは一体何なのかが明確でない。コンピュータに先に述べた原則をのせようとする、「文の前半/後半」とか「新しい情報/古い情報」といった概念を厳密に定義する必要がある。これが表1の計算言語学の特徴として厳密性と書いた一つの理由である。

以上は言語学の理論が厳密性を欠き使いものにならないということを意味しない。人間の直観に基づく理論は理論として有用である。ただそれを計算言語学の立場から見直すことで理論の問題点が明確化し、より精密な理論の構築に向けての努力を促すことができるという点に注目したいと思う。これはむしろ計算言語学の立場から言語学への貢献であると考えられる。実際最近の言語学では、語彙項目の充実とパターンマッチ(別号ユニフィケーション)の重要性が指摘され(表1参照)、これらを巡って新しい言語理論が提案されている[Sells 86][Shieber 86][郡司87]。言語学者と計算言語学者との共同研究も行われるようになってきた。この傾向は今後ますます強まると思われる。

最近筆者の研究室にチョムスキーの統率・束縛理論をコンピュータ上にモデル化し動作させたいとして訪ねてきた言語学科所属の学生がいる。このような学生が現れたこと自体、言語理論が精密になってきた証拠であるのかも知れない。学生に限らず一般に言語学者が同様なことを試みたと

き、最初に出くわす障害はプログラムを書かなければならないということだろう。特に文法規則と辞書を用いて文の合文法性を調べるパーザと呼ぶプログラムを作成しなければならない。幸いにして最近パーザについての障害は取り除かれつつある。優れたパーザが世の中に存在しているので、それを使えばよいからである。図1で言えば言語学者は文法規則と辞書の記述だけを行えばよいことになる[田中 86]。言語学者が自分の言語理論をコンピュータに組み込み動作させることが比較的容易になってきており、本章の[2]で述べたことが現実味を帯びてきたといえる。

言語の解析の概略を図1に示す。

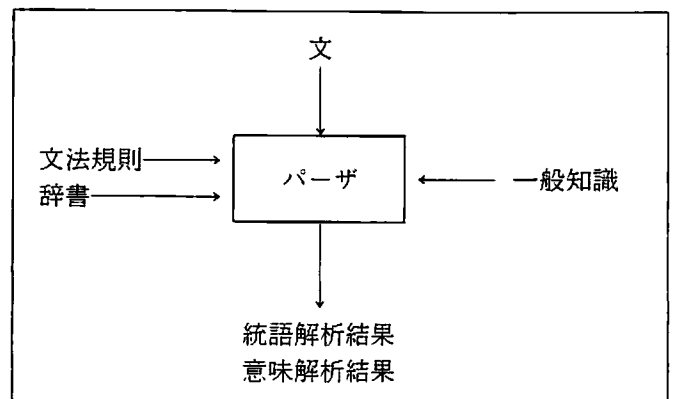


図1 言語の解析

言語学では文法規則と辞書を次のように記述することがある：

- (1) NP → NP P NP
- (2) NP → N
- (3) N → 花
- (4) N → 蜜蜂
- (5) N → 野原
- (6) P → と

ここで各規則は、“→”の左の記号が“→”の右の記号列に書き換え可能なことを意味している。規則を次々に適用し書き換えてゆくことにより、以下のようにして記号 NP から「花 と 蜜蜂」を生成することができる。

- (1)を適用          NP => NP P NP  
 NPに(2)を適用      => N P NP  
 Nに(3)を適用       => 花 P NP

Pに(6)を適用      => 花 と NP  
 NPに(2)を適用    => 花 と N  
 Nに(4)を適用      => 花 と 蜜蜂

上で(3)の代わりに(5)を適用すれば「野原 と 蜜蜂」を得ることができる。このように少なくとも言語学では抽象的な記号(たとえばNP)から言語を作り出す過程(生成過程)に興味があった。規則の矢印の向きが言語を生成する向きと一致しているのはそのためである。これに対して計算言語学では、現実の文が文法にかなっているかどうかを調べることに興味がある。言語を解析する過程に興味があるのである。言語学でも解析に興味がないわけではないが、言語学者は彼の直観で解析結果(統語解析結果)を書き下すことで満足する。ところが計算言語学者は自分の直観による解析だけでは満足しない。なぜなら解析にコンピュータを用いるので解析アルゴリズムを考えなくてはならないからである[野村 88][松本 88]。しかもそれは高速でなければならぬ。さらに次章で述べる曖昧さを解消する問題も解決しなければならぬ。これは計算言語学から心理学への接近を意味する。

### 3. 計算言語学と心理学

一般に意味的な解析を無視し統語解析のみに限定すれば、数十語の単語からなる文を統語解析すると数百から数万の統語解析結果が得られることがある。たとえば「野原の花と野原の蜜蜂」という名詞句に対して、思い付くままに結合の仕方を挙げると5通りある。文法規則は結合の仕方を与えるものであるから、純統語論的にいえば5つの統語解析結果が得られる。

[[野原 の 花]と[野原 の 蜜蜂]  
 [[野原 の [花 と 野原] の 蜜蜂]  
 [[野原 の 花] と 野原] の 蜜蜂  
 [野原 の [[花 と 野原] の 蜜蜂]  
 [野原 の [花 と [野原 の 蜜蜂]]]

このようにして、文が長くなるにつれて曖昧性の数の組み合わせが増え、莫大な数の統語解析結果

が得られることになる。しかもそれらのほとんどが意味的に異常な解釈を強制するものであることも問題である。計算言語学の研究者の悪戦苦闘の歴史は、解析結果に含まれる曖昧性をどのように解消するかにあったと言ってよい。

ところで人間はどのようにして言語を解析し理解しているのだろうか。解析は言語理解の過程で最も重要な部分である。人間の言語理解の過程が解明できれば、それをコンピュータに組み込み解析の効率を向上させることができるかもしれない。言語理解は人間の認知機構の一部である。言語に関する人間の認知機構の解明は心理学の最も魅力的な研究課題である。この問題を巡って計算言語学は心理学に接近することになる。ただしここでいう心理学は旧来の心理学でなく新しい心理学でなければならぬ。新しい心理学というより認知科学の一分野に接近するといった方が正確かも知れない。

さて人間はいかに長い文であっても言語理解の過程で数万もの統語解析結果を得ているとは到底思われぬ。文を読んだり聞き終わった段階で一つの意味解釈を得ていると思われる。言語の解析はほとんど曖昧さなく(ほぼ決定的に)進行しているように見える。そして最後に得られる文の意味解釈には、おそらく一つの統語解析結果が対応している。言語学者が直観に基づき文の統語解析結果を書き下すのは、そのようなものに違いない。

人間が言語をほぼ決定的に解析しているとしたら、本章の前半で述べたように、文法規則を規則的に適用するといった純粹に統語的な解析とは異なる方法をとっていると考えられる。この問題に対して人工知能や認知心理学の研究者の考えを二つだけ示そう。

- (1) 文に含まれる単語を幾つかバッファにため込み(先読みし)統語的な処理を行う[Marcus 80].
- (2) 統語解析と意味解析とが融合した処理を行う。時には文外の情報をも利用する[Wingrad 72].

これらはいずれも人間が言語解析過程で採用して

いる戦略のように思える。計算言語学の立場からはこれはいずれも計算アルゴリズムとして実現しなければならない。まだ解明すべき数多くの研究課題が残されているのである。心理学者の側からの共同研究がこれから必要であると思われる。特に意味解析には問題が山積している。以下に意味解析に関連する問題を思い付くままに列挙してみる。

- (1) 意味解析には知識が必要であるとされているが、我々が言語理解を行うときに用いている知識の構造が一体どのようなものか良く分かっていない [Miller 76].
- (2) 文に明示的に述べられていないことを次の文で受けることが頻繁に起きる。知識を用いて文間を補う推論が必要になるのであるが、そのメカニズムに対して研究がさらに必要である [Schank 77].
- (3) 比喩的な表現を理解するメカニズムの解明がほとんどなされていない [佐藤 78] [Lakoff 83] [山梨 88].

ここで(2)について少し補足しておきたい。たとえば我々は次の文を容易に理解できる。

「落石だ!」と愛一郎は叫んだ。美那子は急いで岩蔭に身を潜めた。

我々はなぜ美那子が岩蔭に隠れたかが分かる。常識を使い、前文から岩蔭に隠れた理由を容易に推論することができる。言語理解の過程を徹底的に追求して行くなれば、いずれこの問題に突き当たるだろう。高度な言語理解システムを作成しようとするなら、この問題を解決する必要がある。

以上は新しい学問である認知科学の研究として、人工知能の研究者や言語学者をも含めた研究をすすめる必要があるだろう。

#### 4. 計算言語学と哲学

言葉を使うということは、言葉の担う情報を相手に伝えることである。言葉の担う情報は直接的な表現をとることもあれば間接的な表現をとることもある。たとえば、相手に対して

「水が飲みたい」

といったとき、相手に伝えたい真の意味は、

「コップに一杯水を持ってきてくれ」

ということであるかも知れない。このことをもって間接的に表現して

「喉が乾いた」

ということもある。言語のこのような側面は発話行為 (speech act) と呼ばれている [Austin 63] [山梨 86].

情報を伝達するという立場から言葉の意味を考察することは、哲学で古くから論じられてきた。これは言語を理解し仕事をしてくれる真に知的なロボットを作る場合に大きな問題になる。上の例から明らかのように、ロボットが文字どおりの意味を理解することだけでは不十分だからである。

哲学者はこれまで信念の構造について議論してきた [Barwise 83] [片桐 88]. 信念は自分に対する信念だけでなく、話し相手の信念のモデルを考慮することが重要であることが分かってきた。話し相手の信念のモデル化は、話し相手の意図の推測や嘘の検出という問題にも関係する。さらには誤解の問題にも関係する [Joshi 81]. 誤解の問題の重要性は次の会話を考えると理解できるだろう。

社長：今日 11 時に星野部長に面会する予定になっていると思うが。

秘書：そのような予定にはなっておりません。11 時に星野部長ではなく藤田部長に会う予定になっているとしたら、秘書の答は相手に誤解を与える可能性がある。なぜなら秘書の応答は、11 時に誰とも会う予定がないと社長に誤解される可能性があるからである。それを察知して、秘書は次のように答えることが望まれる。

秘書：いいえ、星野部長ではなく藤田部長と会う予定になっております。

これまで哲学者の考えることは思弁的であり役に立たないとされてきたように思う。計算言語学のなかの円滑な対話モデルの研究では、哲学者のこれまでの地道な研究成果が本質的に重要な役割を果たすと思われる。この問題に関連して哲学者

の側から計算言語学に興味を持つ研究者が今後多数現れることを期待したい。

参考文献

- [Austin 63] Austine, J. L.: How to do Things with Words, Oxford Univ. Press (1963). 坂本百大 (他訳): 言語と行為, 大修館 (1978).
- [Barwise 83] Barwise, J. and Perry, J.: Situations and Attitudes. The MIT Press (1983).
- [Gardner 85] 佐伯 (他訳): 認知革命, 産業図書 (1987).
- [Grishman 86] Grishman, R.: Computational Linguistics—An Introduction. Cambridge University Press (1986).
- [郡司 87] 郡司隆男: 自然言語の文法理論, 産業図書 (1987).
- [Joshi 81] Joshi, A., Webber, B. and Sag, I.: Elements of Discourse Understanding. Cambridge Univ. Press (1981).
- [片桐 88] 片桐恭弘, 談話の世界 (5章), in [田中 88].
- [Lakoff 83] Lakoff, G. and Johnson, M.: Metaphors and We Live by, The Univ. of Chicago Press (1980). 渡辺昇一 (他訳): レトリックと人生, 大修館 (1986).
- [Marcus 80] Marcus, M.: A Theory of Syntactic Recognition for Natural Language. The MIT Press (1980).
- [松本 88] 松本裕治, 統話解析の手法 (3章), in [田中 88].
- [Miller 76] Miller, G. and Johnson-Laird, P. N.: Language and Perception, Harvard Univ. Press (1976).
- [長尾 83] 長尾 真: 言語工学, 昭晃堂 (1983).
- [野村 88] 野村浩郷, 自然言語処理の基礎技術, 電子情報通信学会編, コロナ社 (1988).
- [佐藤 78] 佐藤信夫: レトリック感覚, 講談社 (1978).
- [Schank 77] Schank, R. C. and Abelson, R. P.: Scripts, Plans, Goals and Understanding, Lawrence Erlbaum Assoc. (1977).
- [Sells 86] Sells, P.: Lectures on Contemporary Syntactic Theories, CSLI, Stanford University (1986). 郡司隆男 (他訳): 現代の文法理論, 産業図書 (1986).
- [Shieber 86] Shieber, S. M.: An Introduction to Unification-Based Approaches to Grammar, CSLI, Stanford Univ. (1986).
- [田中 86] 田中穂積, 新田義彦: ロジックプログラミングと計算言語学, 情報処理, vol. 27, No. 8 (1986).
- [田中 88] 田中穂積, 辻井潤一 (編著): 自然言語理解, オーム社 (1988).
- [Winograd 72] Winograd, T.: Understanding Natural Language, Academic Press (1972). 淵 一博 (他訳): 言語理解の構造, 産業図書 (1977).
- [山梨 86] 山梨正明: 発話行為, 大修館 (1986).
- [山梨 88] 山梨正明: 比喩と理解, 認知科学選書 17, 東京大学出版会 (1988).
- (たなか・ほづみ, 東京工業大学・工学部)

ORに親しみORを役立てる

# オペレーションズ・リサーチ

1989年2月号

## 特集 情報ネットワーク

金融機関の国際ネットワーク  
セキュリティ情報ネットワーク  
企業情報ネットワーク  
組織内情報ネットワークと組織のゆらぎ  
情報ネットワークと意味の自己組織化  
情報ネットワーク化と経営意思決定

1989年1月号

## 特集 投資と金融のOR

理財工学のすすめ  
多目的問題としての資本資産評価モデル  
疑似インデックス・ポートフォリオ  
投資リスクと動的投資理論  
ファイナンスにおける最適停止問題

連載講座  
証券投資技法の基礎と概要(最終回)

### ■日本OR学会入会のご案内■

正会員(個人)	年会費	12,000円
	入会金	1,200円
学生会員(個人)	年会費	5,000円
	入会金	600円
賛助会員(法人)	年会費	A種 95,000円
		B種 48,000円
	入会金	不要

オペレーションズ・リサーチ誌および論文誌が配布されます。

### ■入会手続き■

入会申込みのご希望があれば、会費振込用紙、原簿等の必要書類をお送りいたします。

各号1部 850円 年間購読料 9,600円 (送料含む)  
お申し込みは下記に

(社)日本オペレーションズ・リサーチ学会

113 東京都文京区弥生2-4-16 学会センタービル

電話 03-815-3351