解説

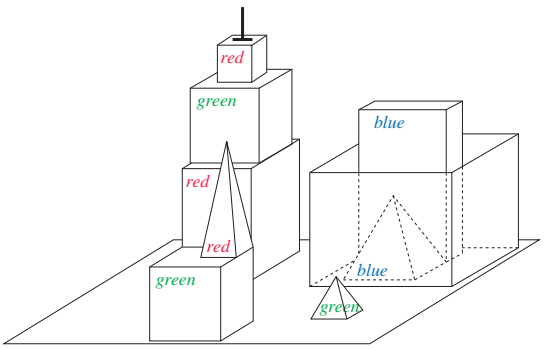
# ロボットにおける言語理解*

徳永健伸（東京工業大学 大学院情報理工学研究科）**・田中穂積（中京大学 情報理工学部）***

## 1. 背　　景

音響技術をロボットに適用しようとする究極の目的は，ロボットに言葉を理解させ，人間との自然なインタラクションを実現することであろう．もちろん，物がぶつかる音や，壊れる音など，我々が耳にする音には，話し言葉に限らず様々なものがあり，たとえば，その音源の位置を正確に同定する技術もロボットにとっても重要である．しかし，言語を使用する能力は人間の知性を特徴付ける重要な要素であり，知的なロボットを実現するためには，その内容を理解し，言語を介した人間とのコミュニケーション能力が不可欠である．

一般に，言語を介して人間とコミュニケーションするシステムは対話システムと呼ばれ，これまでにも多くの対話システムが研究されてきた．ロボットとの対話システムという意味では，1970 年代初頭に開発された Winograd の SHRDLU が最初のものである．SHRDLU は，図 1 に示すように，コンピュータディスプレイ上に表示された様々な色や形の積木が存在する閉じた世界を対象としている．人間がキーボードを介してシステムに自然言語で命令をすると，積木の世界の中にあるロボットアームが命令に応じて動作し，指示どおりに積木の世界の状態を変えてくれる．また，積木の世界の現在や過去の状態に関してシステムに質問をすることもできる．SHRDLU は当時として考えうる自然言語処理のほとんどすべての処理を網羅し，多くの言語現象を扱うことができた最初の対話システムである．また，ディスプレイを備え，世界の変化を動画として確認できるという点でも画期的であった．




**図–1**　SHRDLU の積木の世界 [14]

SHRDLU 以降も多くの対話システムが研究されてきたが，その中心的な対象領域は，対話を通じてデータベース中の情報を提供したり，旅行計画の立案をしたりする，いわゆる情報探索型の対話に移っていった．これらの対象領域では実際に動作して世界の状態を変えることができるロボットを利用する利点は少ない．

また，マルチモーダル対話システム[1]と呼ばれる研究も活発におこなわれている．この研究分野では，情報伝達ために複数のチャネルを効果的に組み合わせることによって，人間とのコミュニケーションをより円滑にしようという試みがなされている．たとえば，システムを擬人化し，表情を持たせたり，視線をうまく制御することによって，人間とのインタラクションをより自然なものにする研究がある．このためには必ずしもハードウェアのロボットを使うわけではなく，コンピューターグラフィクスで生成されたアバタを使うことが多い．このようなアバタは「身体を持つ会話エージェント (embodied conversational agents) [2]」と呼ばれ，近年特に注目を集めている．

* Language understanding by Robots.

** Tokunaga Takenobu (Department of Computer Science, Tokyo Institute of Technology)
*** Tanaka Hozumi (School of Information Science and Technology, Chukyo University)

[1] 音，テキスト，映像，画像などの複数の種類の情報チャネルを通じて対話をおこなうシステム．

## 2. 課　　題

例として部屋の家具の配置を相談している家族の会話を考えてみよう．

**対話例 (1)**

(1) 夫 :このソファは右の壁の方がいいんじゃない?
(2) 妻 :(振り返って) この出窓のあたり?
(3) 夫 :いや，君から見て右．
(4) 妻 :ちょっとここには入らないんじゃないの．
(5) 夫 :(子供に向かって) メジャーをお母さんに取ってあげて．
(6) 子 :どこにあるの？
(7) 夫 :机の引き出しにあるだろ．
(8) 子 :あった．
(9) 夫 :それをお母さんに渡して．

これがロボットの家族の会話であったなら，このロボットには少なくとも以下のような能力が必要となる．以下では，これらの研究課題について詳細に説明する．

- オブジェクトの空間的関係を理解できる能力
- 言語情報とパラ言語情報を統一的に扱う能力
- 漠然性を扱う能力
- 協調作業をする能力

### 2.1 オブジェクトの空間的関係を理解できる能力

前後左右や上下など物体間の空間的な関係とその言語表現の間の関係については哲学や認知科学などの分野で数多くの研究がある．対話例 (1) において，互いに向い合っている状況で「右の壁」(発話 (1)) という場合，話者の視点に立つのか，聴者の視点に立つのかで,「右」の解釈が逆になってしまう．実際に発話 (2) から妻がこの解釈を誤っていることがわかったので，発話 (3) では，視点を明確にして「君から見て右」と言い直している．このような空間的関係の解釈は，単に視点だけの問題ではなく，認知科学では，より一般的な参照枠という概念を使って説明されている．一般に空間関係の表現では基準となるオブジェクト (参照物) を利用する．たとえば,「ソファの右」という表現では，ソファを参照物としてある方向・場所を表現している．このような解釈をするためには視点に加えて，参照物の特徴なども考慮する必要がある．

参照枠を決めるモデルは認知科学の立場からいくつか提案されている．たとえば，Levelt は座標系と参照物がそれぞれ話者であるか話者以外であるかによって，参照枠を 3 種類に分類しているし [4]，Retz-Schmidt は，参照物自身が方向性を持つかどうかという要因と視点からやはり 3 種類に分類している [8]．ここでは，Levinson の分類について説明しよう [7]．Levinson は参照枠を内在参照枠，絶対参照枠，相対参照枠の 3 つに分類している．内在参照枠は，参照物自身が固有の方向軸を持つ場合に，この参照物の固有軸を積極的に利用する場合に設定される．たとえば,「私の右」という表現では,「私」の固有軸を基準に内在参照枠利用して，右手の方向・場所を指すことになる．参照物が固有の方向軸を持たない場合 (たとえば,「木の右」) では，内在参照枠は使うことができない．絶対参照枠は東西南北などのように，参照物や対話参加者とは独立した基準を使う場合に設定される．一方，相対参照枠を使った解釈では，話者と聞き手，の位置関係や参照物の性質などがすべて関連する．対話例 (1) では，夫が妻を参照物とした内在参照枠を使って,「右の壁」と表現したのに対し，妻は夫から見た相対参照枠を使って，これを解釈したために，誤解が生じたと説明できる．

このように認知科学の研究は参照枠を分類し，現象を記述することに主な興味があるが，参照枠をどのように設定すればよいのか，その具体的な手続きについては必ずしも教えてくれない．Herskovits は参照枠の分類よりも，それを決定付ける要因を中心にこの問題を整理し，座標系の原点，軸の順序，軸の方向の 3 つの要因によって参照枠が決定されるとし，最初の 2 つの要因については決定方法を述べているが，肝心の軸の方向の決定については明確な答えを出していない [4]．ロボットに言語理解をさせるためには，参照枠を計算する方法を組み込んだ実験システムを構築し，動作させながらこれらの要因について実証的に明らかにしてゆく必要がある．

### 2.2 言語情報とパラ言語情報を統一的に扱う能力

人間同士の対話の中では，暗黙のうちに多くの情報が言語表現以外の手段によって伝わっている．たとえば，表情や手の動き，視線，あるいは声の

調子などのパラ言語情報[2]は言語で表現された情報を補完する役割を担っている．対話例 (1) でも，単に言語表現だけではなく，言語表現と視線の移動との同期を考慮しないと，発話 (2) や (5) は正しく理解できない．

Cassell ら [2] は話しをする人間の動作をビデオに撮影・分析し，ソフトウェアロボットに実装する試みをおこなっている．また，自然な視線の動きの実現，コンピュータアニメーションにおいて発話と口唇の動きを同期させる口唇同期 (Lip Sync) と呼ばれる技術，表情の生成などは，人工知能における仮想エージェントの研究分野では活発に研究されているテーマである．これらの研究は，コンピュータグラフィクスや音声認識の技術の進歩を前提としており，最近のこれらの研究分野の進展によって始めて可能なったものである．

### 2.3 漠然性を扱う能力

コンピュータによる言語理解においては曖昧性は様々な解析の段階で問題になる．たとえば，前節の対話において，発話 (9) で使われている代名詞の「それ」が「メジャー」を指しているということは，人間ならばすぐにわかるが，コンピュータでこれを同定するのはそれほど容易ではない．テキスト中の既出のオブジェクトは，メジャーの他にも机や引き出しやソファなど「それ」で指されそうなものがいくつもある．これは照応の曖昧性と呼ばれ，これを解決する処理は照応の解消と呼ばれている．

この例は指示対象を対話を書き起したテキストの中に見つけることができるが，その場の状況を考慮しないと解消できない照応もある．このような照応は一般に外界照応と呼ばれる．たとえば，以下のような対話を考えよう．

**対話例 (2)**

客 :(大根を指差して) これ，ください．
八百屋 :一本でいいの？
客 :はい．
八百屋 :じゃ，200 円ね．

最初の客の発話中の「これ」が指すオブジェクトは，その場には存在するがこの対話を書き起したテキスト中には現われない[3]．この例でもわかるように，外界照応は言語を記号の中に閉じた系として考えていたのでは扱えない．

言語処理の研究において言語の曖昧性は中心的な課題であり，多くの研究がおこなわれてきたのに対して，漠然性に関する研究は非常に少ない．上述した照応の曖昧性のように，一般に言語処理における曖昧性の解消は，多くの候補の中から正しいものを選択する離散的な過程としてとらえることができる．これに対して漠然性は数え上げることができない候補からもっともらしい答を見つける連続的な過程であるといえる．たとえば，対話例 (1) の発話 (3) においてある場所を「その辺」という表現で指示しているが，この場合，指示されている場所を数え上げてその中から正解を選ぶという処理は適切ではない．もちろん最終的に「その辺」に物を置く場合には，位置は一意に決まることになるが，対話の中でやりとりされる「その辺」という言語表現は記号的なものでありながら，それが指示する場所には連続的な広がりを許している．別の言い方をするならば，我々が対話をおこなう際に用いる言語表現には，実世界に照らして非常に限られた不完全な記述しかしていないものがある．にもかかわらず，我々はそれを意識せずにコミュニケーションできるのである．言語を理解するロボットとの間で言語や行動を通じてインタラクションしようとすると，このような漠然性を扱うことが不可欠となる．

ロボットの行動計画は古典的な人工知能の分野では記号処理を基礎としたプランニングによっておこなわれてきた.「その辺」という表現を単に記号として扱っている限り，古典的な手法は使えるかもしれない．しかし，この例のように，言語でやりとりをしながら，位置を決め，そこに何かを置く動作までおこなおうとしたら，位置に関する不完全な言語表現から最終的には，具体的に正確な座標を求める必要がある．このためには従来の古典的な記号処理と空間座標のような連続量のギャップを埋める枠組が必要となる．

### 2.4 協調作業をする能力

対話例 (1) の発話 (5) で，夫が子供にむかって「それ (メジャー) をお母さんに渡して」と言って

[2]ここでは，音響的な情報だけでなく，発話にともなうジェスチャ，表情などの非言語的情報も含む．

[3]括弧内はト書きであって，発話を書き起したものではない．

いるが，子供がメジャーを母親に渡すためには，子供が「渡す」動作をするだけではなく，母親の方も同時に「受け取る」動作をしなければならない．このように 2 人が協調して動作をしないと目的は達成できない．この例にはもうひとつ興味深い点がある．発話 (9) は表面的には子供に向けられたものであるが，実際にはその場にいる母親にも聞こえていて，母親に対するメッセージも込められている．これまでの対話システムでは 1 対 1 の対話を扱うものが多いが，このように対話参加者が 3 名以上いるような例では，ひとりの発話が複数の対話参加者に聞こえることがある．このような対話は多人数会話 (multi-party dialogue) と呼ばれ，最近，盛んに研究されるようになっている [13]．多人数会話では，発話が特定のひとりに対するメッセージの場合もあれば，複数，あるいは全員に対するメッセージの場合もある．誰に対するメッセージなのかは状況に依存し，これを判断することが必要となる．

対話例 (1) では，対話を通して複数の人間が協調的に計画を立て，家具を協調して移動しようとしている．協調作業というとこのような物理的な作業を連想しがちだが，実は対話そのものが対話参加者による協調作業となっている．裏を返せば，対話が成立するためには協調作業が必要である．より具体的には，対話が成立するためには，対話参加者が共通の基盤 (common ground) を共有する必要がある．つまり，お互いが考えていることを互いに知っている必要がある．たとえば，対話例 (1) の発話 (2) で，妻は夫が直前の発話で指示している「右の壁の方」がどこであるかを確認している．実際，妻のこの理解は夫の意図した場所とは異なっており，発話 (3) で，夫は妻の誤解を訂正することになる．この例からわかるように，対話の目的は，言語表現のやりとりによって互いに共通に理解した基盤を構築することである．この過程は，基盤化 (grounding) と呼ばれている．基盤化では，共有すべき情報内容そのものを伝達すると同時に，それが正しく伝達された (基盤化された) ことを相手にフィードバックする必要がある．たとえば，対話例 (3) で B の最初の発話は A の最初の発話が聞き取れなかったということを A に伝える機能がある．また，B の 2 番目の発話は，A の誘いに対する肯定的な答えを伝達する機能と同時に，A の 2 番目の発話を正しく理解したことを A に伝える機能も担っている．

**対話例 (3)**

A :明日，映画行かない?
B :何?
A :明日，映画行こうよ．
B :いいよ．

対話例 (1) や (3) からわかるように，対話における基盤化はいつもスムーズにおこなわれるわけではなく，聞き違いや誤解などのさまざまな障害によって妨げられる．これらの例でわかるとおり，人間はこれらの問題を対話によって解決する．このように対話がうまく進行するには基盤化が必要であり，基盤化をうまく進めるために対話をおこなう．ロボットに対話をさせるためには基盤化を扱うメカニズムが必要となる．Traum らは有限状態機械を用いて，基盤化の状態，つまり，何が基盤化され，何が基盤化されていないかを管理する枠組を提案している [12]．

Traum の基盤化の計算モデルは主として (音声) 言語による対話を想定しており，音声情報や言語情報を利用して，基盤化の状態遷移をおこなう．しかしながら，実際に動作可能なロボットに基盤化のモデルを実装するとなると音声・言語情報に加えて，ロボットの動作も考慮する必要がある．たとえば，机の上の本を本棚に戻すために，ロボットに対して「その本を右の本棚に戻して」という発話をしたとしよう．これに対して，ロボットは指を指すなどして,「この本をこの棚に戻すんですね」のように対象となる本や棚を確認することもできるし，確認することなくすぐに命令された動作をおこなうこともできる．前者の場合，話者はロボットが正しく対象を理解しているかどうかを，ロボットの動作と発話から確認して，そのまま受け入れるか，必要なら修正するであろう．いずれにしても，このやりとりにより，動作の対象となる本や棚が基盤化される．一方，ロボットが命令されてすぐに動作を開始する場合，話者はその動作を観察し，自分が意図したとおりにロボットが対象物を理解しているかどうかを見極める必要がある．そして，ロボットの動作の過程でロボットが誤解しているとわかった時点で，誤解を正す発話をするであろう．このように実際に動作するこ

とによって，基盤化されているかどうかの手がかりを与えることができる．ロボットとの対話を考える場合，このような動作による基盤化も考慮して従来のモデルを拡張する必要がある．

## 3. 展　　望

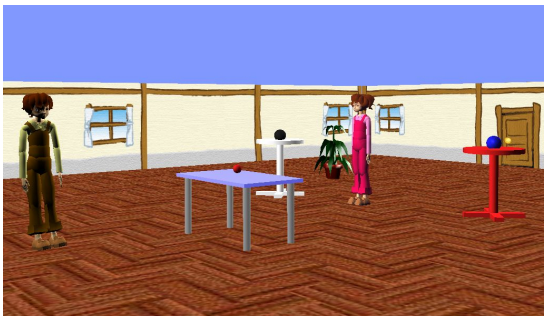
身体性を持ち，言語を通して人間とインタラクションできるソフトウェアロボットの研究が近年注目を集めている．これらは「身体を持つ会話エージェント (embodied conversational agents)」と呼ばれている．これらの研究で重要な点は，単にコンピュータグラフィックスによって精緻なアニメーションを生成するだけではなく，その多くが言語能力を重要視していることである．

このような研究分野は必然的に学際的なものとなる．すぐに思い付く関連分野として，コンピュータグラフィクス，音声言語処理，計算言語学，認知科学，哲学，言語学などがあげられる．著者らのグループでもこれらの関連分野の研究者を組織し，2001 年度から 2005 年まで「言語理解と行動制御」という研究題目で研究をおこなった (文科省科学研究費補助金 学術創成研究 13NP0301)．本節では，一例として我々のプロジェクトを取り上げ，具体的な研究の取り組みについて述べる[4]．

我々のプロジェクトでは，これまで記号の世界に閉じておこなわれてきた言語理解の研究を，実／仮想世界とのインタラクションを導入することによって，状況を考慮した言語理解に発展させることを主な目的としている．特に言語理解の結果として生じるロボットの行動を重要視している．ただし，これは単に言語を解析した結果を視覚化するというだけの意味ではない．

Austin [1] や Searle [9] らの言語行為論に見られるように，言語の使用 (発話) も行為の一種であると考えることができる．逆に発話に対して，物理的な動作や音調などのパラ言語的な手段によって応対することもできることを考えると，ある種の行動は言語の使用と同類であるともいえる．このように人間の活動において言語と行動は密接な関係にあるにもかかわらず，これまで言語処理は言語を閉じた記号系として扱い，ロボティクスでは行動を単なる制御系の問題として扱ってきた．知的なロボットを実現するためには，言語と行動を統一的に扱う必要があると我々は考えている．

この目的を達成するために，我々は仮想世界中に存在する複数のソフトウェアロボットと音声対話によってインタラクションできるプロトタイプシステムを作成し，これをテストベッドとしていくつかの研究テーマに取り組んできた．



**図–2**　プロトタイプシステムのスクリーン

図 2 はプロトタイプシステムのスクリーンショットである．この図では，仮想空間中に 2 体のアバタと机やボールが置かれている．人間は音声入力によってロボットに指令を出し，ロボットはそれにしたがって世界の状態を変化させる．ロボットの行動と世界の変化の様子はアニメーションによって人間に提示される [3]．

このプロトタイプシステムを使って以下の項目について研究をおこなってきた．

**【音声入力における言い直しや言い誤りを扱うための言語処理】** [6]

話し言葉に頻繁に現れる助詞落ち，倒置，自己修復などは，音声対話を困難にする大きな要因のひとつである．これらの現象が複合して現われる現象に対応するために，発話断片の統語的・意味的な性質の類似性を利用して解析する手法を開発している．

**【プランを利用した参照表現の解釈】** [15]

言語解析の研究において，参照表現の指示対象を同定する研究はさまざまな観点から研究されている．我々が扱っている対象領域では，言語表現として現われる表層的な文脈情報より，発話によって話者が達成したい目標やそのためのプランの情報を利用することによって参照表現を解釈する方がより正確な解釈が可能であることを示し，そのための手法を提案している．

[4] http://www.cl.cs.titech.ac.jp/sinpro

**【空間的な漠然性の表現と行動計画】** [11]

すでに述べたように古典的な人工知能の行動計画では，空間の位置を表現するのに記号が使われてきた．しかし，位置を指示する言語表現は漠然としており，それによって指示される位置もある程度の広がりが許される．このような位置の漠然性を古典的な行動計画の手法で扱うために，我々のシステムでは記号表現と位置のもっともらしさを表わすポテンシャル関数を組み合わせたオブジェクトを使っている．これによって，部分的に空間的表現の漠然性を扱えることを明らかにした．

**【構成的な動作の辞書の構成】** [10]

アニメーションを生成するためには，ロボットの動作を定義した辞書が必要となる．しかし，すべての動作を定義することは不可能なので，基本的な動作を定義して，その他の動作は基本動作から構成的に作り出すような機構が必要である．基本動作をどのように定義するか，あるいはそもそも基本動作なるものが定義できるのかについては哲学の分野でも長い論争がある．基本動作を一般的に定義することは困難なので，我々は工学的な立場に立ち，ロボットとの対話により室内の物体を移動するタスクに限定して動作の辞書の構築している．基本的な動作の元となるデータはモーションキャプチャシステムを利用して採取し，これを変換して辞書を構築するアプローチを取っている．

**【修正発話における修正対象の同定】** [5]

前節で述べたとおり，対話を進めるには対話参加者の間の基盤化が必要である．修正発話は相手の誤解を正し，基盤化を進めるための手段のひとつである．動作をおこなうロボットとの対話においては，基盤化の手がかりとしてロボットの動作も考慮する必要がある．従来の基盤化のモデルを拡張し，言語と動作を同時に扱う基盤化のモデルを構築した．

以上，本稿では，ロボットとの対話という観点から次世代のロボットに求めれられる機能について概観してきた．言語を理解し，適切に行動できるロボットが実用になれば，手話通訳や介護サービスなどに応用できるだろう．また，最近のビデオゲームの一部には音声入力をインタフェースとして，ゲーム中のキャラクタを制御するものも出始めているが，キーワードのみを認識して反応する非常に初歩的なものにすぎない．言語を理解するロボットはビデオゲームなどのエンターテイメントにも応用できよう．

## 文　献


[1] J. L. Austin. *How to Do Things With Words*. Harvard University Press, 1962.

[2] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors. *Embodied Conversational Agents*. The MIT Press, 2000.

[3] Kotaro Funakoshi, Takenobu Tokunaga, and Hozumi Tanaka. Conversational animated agent system k3. In *Proceedings of International Conference on Intelligent User Interfaces (IUI 2006)*, 2006.

[4] A. Herskovits. *Language and Spatial Cognition. An Interdisciplinary Study of the Prepositions in English*. Cambridge University Press, 1986.

[5] Funakoshi Kotaro and Tokunaga Takenobu. Identifying repair targets in action control dialogue. In *Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL 2006)*, pages 177–184, 2006.

[6] Funakoshi Kotaro, Tokunaga Takenobu, and Tanaka Hozumi. Processing Japanese self-correction in speech dialog systems. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 287–293, 2002.

[7] S. C. Levinson. *Space in Language and Cognition*. Cambridge University Press, 2003.

[8] G. Retsz-Schmidt. Various views on spatial prepositions. *AI Magazine*, 9(2):95–105, 1988.

[9] J. R. Searle. *Speech Acts. An Essay in the Philosophy of Language*. Cambridge University Press, 1969.

[10] Tokunaga Takenobu., Okumura Manabu, Saitô Suguru., and Tanaka Hozumi. Constructing a lexicon of action. In *the 3rd International Conference on Language Resources and Evaluation (LREC 2003)*, pages 172–175, 2002.

[11] Tokunaga Takenobu, Koyama Tomofumi, Saito Suguru, and Okumura Manabu. Bridging the gap between language and action. In *Intelligent Virtual Agent - 4th International Workshop IVA 2003*, LNAI 2792, pages 127–135. Springer, 2003.

[12] David Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, 1994.

[13] David Traum. Issues in multi-party dialogues. In F. Dignum, editor, *Advances in Agent Communication*, LNAI2922, pages 201–211. Springer-Verlag, 2004.

[14] T. Winograd. *Understanding Natural Language*. Academic Press, 1972.

[15] 徳永健伸, 関谷幸恵, and 田中穂積. 対話システムにおけるプランベースの照応解析. In **情報処理学会第** *65* **回全国大会論文講演論文集**, 2003.