Vol. 33 No. 7　　情　報　処　理　　July 1992

解　説

自然言語処理技術の最近の動向

頑健な自然言語処理へのアプローチ†

松　本　裕　治††

1. はじめに

自然言語処理において培われてきた多くの技術は，システムがもつ辞書や文法によって定義される適格な文（well-formed sentence）を対象とするものである．しかし，実用的な自然言語処理システムは，システムの予測を超えるような入力にも対処できる能力をもつ必要がある．

ここでは，頑健な自然言語処理（robust natural language processing）の要件とそのアプローチについて解説する．頑健な自然言語処理システムとは，多様な言語現象に対処する能力をもち，かつ，文法的に不適格な文に対してもなんらかの結果を返す能力のあるシステムと捉える．広義の頑健な自然言語処理システムは，これに加えて，曖昧な文に対して最も適切な解析結果を返す能力をももつシステムである．曖昧性の問題については，本特集の長尾・丸山両氏の解説[44]で詳しく述べられている．曖昧性の問題をここで述べる問題と完全に切り離して考えることはできないが，本稿では，主として狭義の頑健性に焦点を当てる．

これまでに提案されたほとんどの文法理論や自然言語解析手法は，適格な文を対象としている．一方，並列句や慣用表現のように人間にとっては適格であるが，文法規則として記述するのが困難な言語現象もある．また，われわれが日常的に目にする文においても，英語の人称や数の不一致などのように頻出する誤りや，堅い文法規則では捉えられない会話的な文が現実に存在する．

タイプ入力を用いた自然言語インタフェースにおいて，時制や人称などの文法的誤りと綴間違いや未知語などの語彙レベルの誤りがいずれも10%以上の文にみられるという報告がある[9),62)]．

人間が困難を感じずに理解可能な文の集合と自然言語システムが受け付けることのできる文の集合などの関係を図-1によって説明することができる☆．「システムにとっての適格文」の集合以外は，集合として定義することが不可能であり，この図は決して正確ではなく単なる概念図である．「文法的適格文（図の濃い色の円）」および「人間にとって理解できる文（薄い色の円）」の境界はいずれも人間によって判断される文の集まりであり，個人によってもそれが使用される文脈によっても適格性が左右される．システムにとっての適格文の範囲を前者（濃い色の円）に過不足なく収めようとする努力が，従来の硬い自然言語処理システムであり，頑健なシステムが目指すのは後者（薄い色の円）を過不足なく理解できるシステムであるといえる．

頑健な自然言語処理システムの能力としてあげた次の2点はこれらのそれぞれの努力に対応している．

1. 多様な言語現象に対処する能力
2. 不適格な文に対してもなんらかの結果を返す能力

すなわち，システムの辞書や文法を充実させ，かつ，慣用句や特殊な言語構造の記述とそれを解析するためのメカニズムを実現する努力が「相対的な不適格文」の集合をなるべく多く覆い「システムが間違って解釈する文」の集合を少なくする研究に対応する．また，「絶対的な不適格文」を処

† Approaches to Robust Natural Language Processing by Yuji MATSUMOTO (Department of Electrical Engineering II, Faculty of Engineering, Kyoto University).

†† 京都大学工学部電気工学第二学科

☆ ここで用いている絶対的および相対的不適格性（absolute and relative ill-formedness）という言い方は Weischedel[60]らの用法に準じている．つまり，回復すべき不適格性のうち，人間にとって文法的でないと判断できる文を絶対的不適格文と呼び，本来は文法的に正しいにも係わらずシステムにとって理解できない文を相対的不適格文と呼ぶ．Fass[12]らは言語的意味的な制約を真偽値で捉えることを絶対的，優先度として連続的に捉える考え方に相対的という用語を用いている．

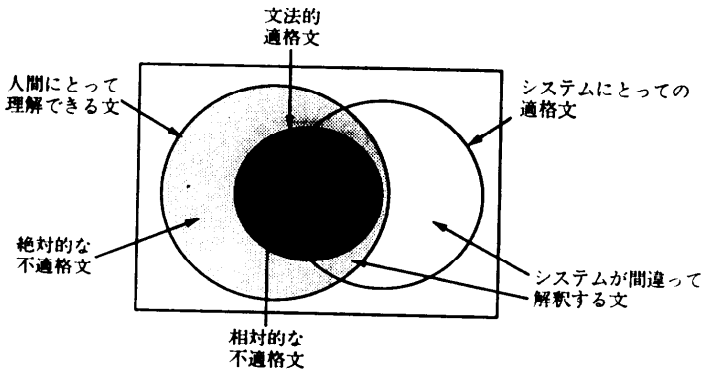

図-1 人間とシステムにとっての適格文と不適格文

理し，「人間にとって理解できる文」を人間と同じように理解できる方法を考案することが不適格文の解析の研究である．

以下の章で，自然言語のさまざまな現象とそれに対する対処の仕方について解説し，頑健な自然言語処理システムに必要な要件を示していくことにする．それぞれのアプローチにはそれぞれの前提や対象があり，それを明確にしておく必要がある．特に注意しておきたいのは次の点である．

対象領域 限定された領域を対象としているか，限定されない領域か．自然言語インタフェースなどは，ある特定の目的に用いることが前提であり，利用者も決まった動機をもってシステムに相対する．このようなシステムでは領域を限定しないシステムとは根本的に異なるシステム設計を行うことになる．

目的 領域を限定しない場合でも，文章中の表面的なキーワードを抽出するためのシステムと，個々の文に対して高い精度の解析を要求されるシステムとでは設計の思想が異なる．

理解の程度 文の真の理解を目指した手法を目指すか，当面の目的に対して最も精度の高い手法を目指すか．統計的な手法によって解析精度をあげる研究などはどちらかというと後者の立場である．

また，現実的な自然言語を扱う頑健なシステムを構築するための要件として，考えておかなければならないのは次の点であると考える．

モジュール性 システムの記述と設計に必要な要件として，語彙や規則，処理システムのモジュール性が高くなければならない．

単調性 個々のモジュールの結合が整合的である必要がある．理想的には個々のモジュールの単純な追加もしくは簡単なインタフェースによって全体のシステムが整合的に構築されることが望ましい．また，システムの量的な拡大によって破綻を来すようなことがあってはならない．

学習的もしくは漸進的なシステム 言語現象や語彙の記述を人手で行うのは大変な労力が必要で，また全体の整合性を保証するのがきわめて困難である．現存の辞書やコーパスを対象にして学習的なアプローチで自然言語処理用の辞書や解析規則を順次獲得するといった方法で量や優先度の問題に対処できる必要がある．

本解説では，続く二つの章において多様な言語現象の取り扱いと不適格文の処理における問題点と解法へのアプローチについて説明する．また，自然言語処理は膨大な語彙ときわめて多様な言語現象を対象とする必要があるため，いくら優れた解析手法や知識の表現法を考案しても，現実的なシステムを実現するためにはこのような量の問題への対処も考えなければならない．4.では，自然言語処理用の辞書や知識の表現に関する問題，および，その自動構築に関する研究について述べる．

2. 多様な言語現象とその解析法

自然言語の解析は，厳密な文法規則を設けてそれに従って解析を行うか，語彙自体に言語的な性質が記述されているかによって大別できる．また，記述自身が特定の処理手続きを仮定しているかという観点でも分類することができる．ATNG（または，ATN. Augmented Transition Network Grammar の略）[72] は拡張された状態遷移ネットワークで記述された文法規則に基づいており，下降型の処理手続きが仮定されている．DCG[48]を初めとする論理文法，FUG[31]，LFG[28]，GPSG[15]などの単一化文法，また LINGOL[52),61)]などは拡張された文脈自由文法に基づいている．これらの文法記述形式の多くは処理手続きとは独立である．

一方，予測駆動解析[57]や WEP (Word Expert Parser)[59]などは個々の単語に構文的，意味的な記

述を行い，明示的な文法規則なしに自然言語文の解析を行うことができる．これらの方法は，前者に比べ，ある意味では，より頑健であり各語の曖昧性や語にまつわる熟語のような特異な構造を柔軟に記述できる．しかし，この方向のアプローチは限定された領域の限られた語彙にしか適用されていない．各語の記述をいかに行うかということ，語彙の増加にともなう計算量の増大，並列構造や痕跡などの広い範囲に影響が生じる言語現象への対処，などが未解決だからである．

本稿では，それぞれの解析法の詳細や得失については論ぜず，どの解析法にとっても問題になる次の言語現象について考える．これらはいずれもきわめて多種多様な構造をともなう現象であり，あらゆる可能な構造を文法規則によって記述するのは非現実的である．

- 等位構造，並列構造 (coordinate structure, conjunction)
- 痕跡，移動 (gap, extraposition)
- 熟語，慣用句 (idiom)
- 挿入句，同格，副詞句 (parenthetic phrase, apposition, adverbial phrase)
- 比喩，換喩 (metaphor, metonymy)

等位構造は，名詞句と名詞句のように同じ構文的な性質をもつ句や節同士を結び付けるだけでなく，一方の句または節からある要素が欠落する場合がある．前者を構成素等位構造 (constituent coordinate structure)，後者を痕跡をともなう等位構造 (coordinate structure with gaps) と呼ぶ．後者の例としてよくあげられる例に次のようなものがある．(痕跡が生じている部分を ϕ によって示す．)

1. *John enjoyed* ϕ and *my friend liked the play.*
2. The car *drove through* ϕ and *completely demolished the window.*
3. I *gave the boy a nickel* and ϕ *the girl a dime.*
4. We *use SPARCstations* and *they* ϕ *MACINTOSHes.*

このような多様性をともなう現象をできるだけ統一的な枠組で対処することが望ましい．等位構造を正しく解析するためには意味的な要因も無視するわけにはいかない．たとえば，上の例の 4. と似た次の例は，SPARCstationとApple MACINTOSH の構成素等位構造と考えるのが自然である．

4′. We use *SPARCstations and Apple MACINTOSHes.*

また，次の 5. は John と Mary それぞれが自分の MACINTOSH をもっていると分散的に読む (distributive reading)のが自然なのに対し，6. では両人が一台のCM-2を共有していると集合的に読む (collective reading) のが自然である．

5. John and Mary use a MACINTOSH.
6. John and Mary use a CM-2.

等位構造の解析法は早くは Woods の ATN にもみられる[73]．これは等位接続詞を発見するとパーザがその前後の語列の中から並置可能な構造を探索するという直接的な方法である．同様の考え方が，Dahl ら[8]によって Prolog 上でも実現されている．また，等位構造をメタ文法規則によって表現し，それを展開し等位構造を扱う具体的な文法規則を生成してそれを用いて解析を行う方法[21]，メタ規則がインタプリティブに痕跡の可能性を考慮しながら解析を行う方法[58]が提案されている．Fong[14]の方法は，等位接続詞の前後を並置することによって等位されている部分列を発見するという意味では，Woods や Dahl らの方法に似ているが，それを宣言的な記述によって行うことにより，等位構造の解析にも生成にも使えることを示した．

日本語の等位構造は，必ずしも and や or のような等位接続詞をともなわないので，より解析が困難な場合がある．Mu プロジェクトで用いられた解析システムでは，等位構造のキーとなるさまざまな表現によって起動される発見的な規則を用意することによって等位構造のスコープを決定している[45]．また，黒橋ら[34]は語の意味的な近さを指標にし DP マッチングの手法を使うことにより，この考え方を拡張している．

痕跡については，ATN で関係節を扱うために Hold と呼ばれるスタックにいずれ痕跡を埋める先行詞を保持し，痕跡が予測されたときにそれをポップアップする手法が提案されている[73]．Pereira[49]や今野ら[33]が同様のアイデアを Prolog 上のパーザへの拡張として実現している．徳永ら[63]は，支配制約という概念を導入して等位構造と痕跡の複合現象を解析する方法を示している．われわれは，チャート法を拡張することによっ

て，痕跡の取り扱いを効率良く行う方法を示している[17]．単一化文法では，SLASH や REL などの素性を用いて，痕跡をともなう長距離依存を宣言的に取り扱っている[51]．

熟語とは，複数の語や句が特別な順序または構造として現われ，字義どおりではない意味をもつものと考えられる．熟語には，完全に固定された語列 (e.g. at any rate, 無用の長物) もあれば，句や節をともなうもの (e.g. take NP into account, 決して〜ない) もある．また，受身や埋め込み文によって変形が起こるもの (e.g. It is taken into account)，それが許されないもの (e.g. *The bucket is kicked, 気がつく，*ついた気) もある．このようにさまざまの熟語がどの言語にも存在する．これらを個別に記述して特徴を明らかにするのは，可能ではあるが，統一的に処理するのは容易ではない．

熟語の解析の手法として，前処理として熟語の可能性を探しておくという方法も考えられるが，効率的にも能力的にも問題である．熟語を文法規則として記述してしまう方法 (たとえば，文献16)) もあるが，これも非現実的である．したがって，熟語の記述は語彙レベルに行う方法が多く取られる[41),59),70),71),74)]．また，熟語を下位範疇化として捉える考え方[10),28)]や，熟語処理をバックグラウンド処理として捉える考え方もある[60)]．

熟語処理は，他の言語現象の解析と同様に，いかに一般的な解析の効率に影響を与えないで処理するか，という問題以外に，字義的な解釈との優先度をどう決定するかということ，また，膨大な熟語の記述をどのようにして行うことができるかという問題をもっている．

挿入句は，文中のある位置に現われ，現在の文または文中の要素に対する補足的な説明を行うために用いられる．**副詞句**についても，役割は異なるが現象的には同様である．このような句や節を文法規則ですべて事前に記述することは不可能である．この問題は，アドホックな処理によって解決できるため，ほとんど注目されていないが，頑健な処理を行うには避けては通れない問題である．

比喩は，「〜のような」などそれを明示する表現があれば，言語の解析の問題というよりも意味や知識の問題である．一方，人を指すのにその所属名を使ったりする**換喩**は，言語現象上では意味的制約の違反となるので，次の章で述べる不適格文の処理の一部として位置付けることができる．Fauconnier[11)]は，彼のメンタルスペース理論で換喩が用いられるときの内的構造を説明している．Lakoff ら[36)]は，換喩には少なくとも次の7種類の関係が考えられるとしている．1)部分と全体，2)生産物とその生産者，3)物とその利用者，4)制御される物と制御する者，5)組織とその一員，6)場所とそこに存在する施設，7)場所とそこで起こる出来ごと．

3. 不適格文とその解析

1.で，不適格文を絶対的な不適格文と相対的な不適格文に分類した．前者は，人間が自分の知っている文法に照らし合わせてなんらかの誤りを感じる文であり，後者は具体的な自然言語処理システムにとって正しいと判断できない文のことである．ここで注意していただきたいが，本稿で扱う不適格文とは無意味な文ということではなく，なんらかの間違いがあるにしても人間には (その文脈を含めて) 理解可能な文のことと仮定する．ここでいう相対的な不適格文は人間にとっては完全な適格文である．

自然言語処理の過程は，形態素解析，構文解析，意味解析，語用論解析に分けて考えられることが多い*．現実の不適格文に現われる現象を，これらの処理レベルによって分類してみると表-1のようになる．相対的不適格文は自然言語処理システムの処理能力に依存するので，ここであげたのは，典型的な例である．前章で述べた複雑な言語現象は構文レベルの相対的不適格性である「システムの能力を超えた構文構造」に分類される．

表中の用語のうち分かりにくいものについて説明する．中止文とは文末が不完全なまま終わる文，言い直しとは文の一部が繰り返されている文で，会話には頻繁に現われる．連続文とは追い込み文 (run-on sentence) とも呼ばれ，句点などが忘れられて二文が一つになってしまった文のことである．特殊構造にはさまざまなものがあり，箇条書きなどの表のような記述，章や節番号，数式など多くのシステムでは考慮に入れられていない

* このような分類は必ずしもこれらの処理が個別に行われることを意味しない．記述の分離と処理の分離を混同してはならない．

表-1 不適格性の分類

	形態素レベル	構文レベル	意味レベル	語用論レベル
絶対的不適格性	綴誤り タイプ誤り 区切り誤り	数・人称の不一致 語順誤り 中止文，語の欠落 冗長文（言い直し） 連続文	必須格の欠落 選択制約違反 換喩	協調の原則違反[40] 照応のための情報不足 電報文
相対的不適格性	未知語 省略表現	システムの能力を超えた構文構造 特殊構造（箇条書きなど）	語彙知識不足 世界知識不足	（文脈的要因による）断片文

が，現実の文章には多く存在する構造である．協調の原則については，文献40）を参照のこと．

不適格文を処理することの重要性は多くの研究者から指摘され[9),18),62),68)]，その解決の重要性が認識されていた．たとえ人間の推敲や校正を経た文章であっても，不適格文の存在は否定することはできない．

不適格文を処理するアプローチとして，不適格文の構造を予測し，適格文を記述するのと同じ方法でそれを記述するものもあるが，処理に無駄があり大きなシステムの実現は困難である．不適格文の可能性は与えられた文が適格文として捉えられない場合にのみ考えればよい．ただし，文が曖昧性をともない複数の解釈の可能性を残す場合には，そのうちのどの解釈が最も妥当な解釈であるかを判断する必要がある．不適格性や換喩を優先意味論の枠組で扱った研究が Wilks らによって行われている[12)]．

不適格文の処理方法にはさまざまなものが考えられているが，多くの方法ではどの程度の不適格文を処理する能力があるかが明確でないことが多い．IBM の Jensen らは，ともかくすべての文に対してなんらかの結果を返す Fitted Parse という考え方を提案している[23),24)]．適格文のみを記述した核文法（core grammar）を用いて文を上昇型並列に解析し，解析が成功しなかった場合，それまでに得られた構造の中から一番もっともらしい構造の列を選び，それをその後の処理の入力とする．この操作が fitting 処理と呼ばれる．システムに与えられる文は次のように特殊な構造をしている場合や断片の場合もあり得る．

1. Example: 75 percent of $250.00 is $187.50.
2. Good luck and good selling.

1. の場合は，“Example” という名詞句と “:” という区切り記号の後に文が存在すると解釈される．また，2. は，全体が名詞句からなる断片であると解釈される．

Fitted Parse では，最初に主構成素（head constituent）を次の優先順位で求める．

1. 主語をともなう時制のある動詞句
2. 主語がなく時制のある動詞句
3. 動詞以外の句（名詞句，前置詞句など）
4. 不定形の動詞句

これによって入力文全体を覆えない場合は，次の優先順位で残りの構成素を主構成素に付けていく．

1. 動詞句以外の構成素
2. 時制のない動詞句
3. 時制のある動詞句

この方法の特徴は，解析システムを引き続く処理の前処理として位置付け，上のようなヒューリスティックスを使って，ともかくなんらかのもっともらしい結果を与えることであり，彼女らが “100% Parsing Capability” と呼ぶゆえんである．

その他のアプローチのほとんどは，文の不適格性の原因を推定し，それが犯している制約を緩和（relaxation）することによってその原因の同定と修正を行う方法を取る．したがって，適格文を対象とした文法によって文の解析を行い，それが失敗した段階で，不適格文の同定の処理が行われる．

Hayes と Carbonell[3)~5),18)] は，領域の限定された対話文を対象として，さまざまなレベルの不適格性に対処するための要件を考察している．不適格性に対しては制約の緩和を行う手法を前提としており，その枠組の中で次のような要件を解析システムに求めている．

1. 解析システムはできるだけ解釈的（interpretive）に動作すること
2. 意味情報の操作が容易にできること
3. 言語の不均一性を考慮していること

4. 下降型，上昇型両方の解析メカニズムをサポートしていること

3番めの要件について補足しておく．言語の記述が文法規則で行われるのであれ，パターンによって行われるのであれ，その記述に含まれる要素はその記述に対して同等の重要性をもっているのではない．不適格性からの回復には，規則内の要素の重要性の差異の情報が有用であることが多く，その差が利用できることが望ましい．

Kwansy ら[35]および Weischedel ら[69]は，Woods の ATN に緩和手法を導入して不適格文の処理を実現している．ここでは，より一般的な後者の方法を説明する．

ATN によって通常の文法が記述され，不適格文を扱う規則はメタ規則と呼ばれる．メタ規則は，ATN の各状態に関する条件の列をもち，それらが満足されたときに実行すべき動作を記述したプロダクション規則である．ATN による通常の解析が失敗すると，メタ規則が呼び出される．メタ規則は ATN が用いていた適格文のための文法のさまざまな制約を緩和するための規則である．メタ規則には，この章の最初で分類したほとんどすべての不適格性が考慮に入れられている．たとえば，意味的な不適格性については，格フレームの必須格条件や格フィラへの制約などの緩和，意味階層上でのより一般的な概念への緩和などが許されている．数や人称などの不一致や換喩的な表現のために意味的制約が犯されている場合などを同じ枠組で記述することができる．

メタ規則として不適格文の処理を統一的に取り扱ったことは優れているが，ATN を対象にしているため，下降型かつ左から右への解析を行うことになる．解析が失敗したときにどの部分に対してメタ規則を適用すればよいかが必ずしも明確ではない．また，ATN は後戻りをともなうアルゴリズムによって動くため，同じメタ規則が繰り返して適用されてしまうことがあり得る．

このような欠点に注目し，Mellish[42]は，チャート法[30]を拡張した不適格文処理の手法を示した．ただし，彼の方法が扱っているのは，語の欠落と冗長な追加という構文レベルの不適格性だけである．基本的には上昇型チャート法によって文の解析が行われ，それが成功しない場合に下降型の規則適用と語彙カテゴリの挿入や削除が行われる．加藤[29]は一般の上昇型解析が失敗した場合に，なるべく多くの部分解析結果が上昇型処理によって得られるように，文法規則の左隅以外のカテゴリからの活性弧生成や活性弧同士の結合を行うことを提案している．得られる情報の増加によって誤りの予測精度は向上するが，同じ文法規則を重複して展開する可能性がある．

自然言語インタフェースでは，直前の文との類似や予測によって，断片的な質問が与えられることが多い．たとえば，列車の切符の予約を行う場合，次のような会話が自然に行われる．

客(1)　来週の日曜の最終で東京までお願いします．

係員(2)　その列車は満員です．

客(3)　では，月曜の始発で．

この対話における(3)の発話はこの文脈からは容易に理解できるが，この文単独では理解が困難である．自然言語インタフェースでは，このように文脈に依存した断片文が多くみられる．

たとえば，LIFER[19]や LADDER[55]では，断片的な発話を前の完全な文の解析木と比較し，その部分解析木とマッチさせることによって断片文の中の省略された語の回復を行っている．Fink ら[13]は，対話の履歴を保存し，新しい発話とそれまでの文との類似性を判断しながら対話の流れを発見する方法を示している．基本的なアイデアはどれも同じである．以前の文とのマッチングや類似性の判定に不確定さが残る．DYPAR-Ⅱ[5]では，格フレーム情報を用いて意味的な整合性の情報も利用している．Lavelli ら[37]は，同じ考え方をチャート構文解析に従ってアルゴリズム的に記述している．断片文の問題は意味的な影響がより重要であり，構文主導の方法の利点は少ないのではないだろうか．

4. 辞書およびコーパスからの言語知識抽出

たとえ対象領域が限定された自然言語処理システムであっても，使用される文型に制限を与えるなどしなければ，大規模な語彙と多様な言語表現に対応することはできない．自然言語理解のための大規模な辞書を構築すること，すなわち，言語に含まれる大量の単語それぞれがもっていなければならない構文的，意味的，語用論的（ある場合には形態素的，また，音韻的）情報を記述し蓄積

することは，頑健な自然言語処理システムを実現する上できわめて重要である．最近，特に，この目的をもつ研究が多く行われるようになってきた．

自然言語処理研究にとっては，たとえ表面的なものであっても大量語彙の辞書が電子化されることが重要である．十分な語彙を包含していなければ現実の文に対処することができない．この意味で，日本電子化辞書(EDR)[47]のようなプロジェクトのもつ意味は大きい．また，情報処理振興事業協会(IPA)は，IPALという計算機可読な日本語の基本動詞辞書[26]および基本形容詞辞書[27]を作成しており，各単語の語義の分類と格フレームの付与を行っている．

電子化された大量の言語テキストデータをもつことも自然言語システムの性能を向上する上で重要である．テキストデータは，なるべく多くの異なる分野から多くの文を集めることが望ましい．Brown Corpusはその代表的なものであるが，最近では，ACL (The Association for Computational Linguistics) の Data Collection Initiative (DCI)*，ACH (Association of Computes and the Humanities), ACL, ALLC (The Association for Literary and Linguistic Computing) の Text Encoding Initiative (TEI)**のように機械可読の大規模テキストのための共通フォーマットを設定しようという動きもある．ヨーロッパでは，European Corpus Initiative (ECI)***といった多言語コーパスの共有化の動きがある．また，構文的に依存構造解析され，照応関係のリンクが張られたテキストデータを蓄積する Textual Knowledge Bank (TKB) というプロジェクトも行われている[56]****．

このほか，いくつもの研究機関で大規模なコーパスの整備が行われている．大規模なテキストコーパスが利用可能になったことを背景にして，それらを統計的に解析してさまざまな言語データを得ようという研究が最近数多く行われている．詳しくは最近の ACL や COLING の論文集を参照されたい．

大規模コーパスや機械可読辞書から語彙に関する構文的意味的情報を抽出する研究も多くみられるようになってきた．たとえば，鶴丸ら[65]は，国語辞典の記述から名詞間の階層関係を自動的に抽出する研究を行っている．人間用の辞書の記述はこのような目的のためにはきわめて情報が希薄であり，有用な情報を得るのがむずかしいが，この方向の研究は重要である．冨浦ら[64]は，IPALの基本動詞辞書の語義文から動詞間の階層関係を抽出する研究を行っている．また，コーパスから語義を抽出する研究が Church[6]，Hindle[20]，Brent[2]らによって行われているが，統計的手法を用いるか，簡単な文解析を行うに留まっている．この理由は，語彙に関する知識を得るために単語のさまざまな用例を現実の文から得ようとすると，それらを正確に解析するために語彙に関する詳細な辞書が必要であるという堂々巡りに陥るためである．語彙に関する知識が希薄な限り，ほとんどの文の構文的もしくは意味的な曖昧性が解消できず，正確な言語データを得ることが困難となる．

二言語以上の対訳コーパスを用いれば，複数の語義をもつ単語の意味的な曖昧性やさまざまな用法がより詳細に抽出できるかも知れない．Klavans[32]や Dagan[7]らは，この方向の研究を行っているが，動詞の異なる用法や曖昧な語義の解消などを表層的な処理によってのみ行っている．

われわれのグループでは，英語と日本語の対訳テキストを構文解析し，両言語の依存構造間の単一化という概念を導入して，単言語だけの解析では解消できない曖昧性の解消が可能であることを示した[66]．和英辞典に現われる例文について，日本語文単独では10数％の文に対してしか曖昧性のない解析結果が得られないにもかかわらず，両言語の結果を単一化することにより，約60％の文に対して曖昧性のない解析結果が得られた．また，こうして得られた解析済みの結果から動詞の格フレームの抽出を行っている．このとき，両言語の文の解析結果が一つにまとめられていることにより，ある日本語の動詞がどのような英語の動詞に翻訳されているかをみることにより語義的な曖昧性解消の手掛かりが得られること，また，日本語の一つの格助詞が英語の複数の異なる格や前置詞を用いて訳されていることなどから格助詞の複数の用法の手掛かりが得られる．このようなことから日本語の動詞の格フレームがかなり自動的

* 問い合わせ先: Mark Liberman, Department of Linguistics, University of Pennsylvania (myl@unagi.cis.upenn.edu)
** 問い合わせ先: L. Burnard, Oxford University
*** 問い合わせ先: Henry Thompson, University of Edinburgh (eucorp@cogsci.ed.ac.uk)
**** 問い合わせ先: Victor Sadler, BSO/Language Technology, Utrecht (sadler@bsolt.uucp)

表-2 獲得された「書く」の格フレーム

格フレーム1

格ラベル	意味カテゴリ	格要素の例
〈*obj*, を〉	13122(通信)	手 紙
〈*subj*, が〉	12000(われ・かれ)	わたし, あなた
〈*to*, に〉	12xxx(人間活動の主体)	友達, 母

格フレーム2

格ラベル	意味カテゴリ	格要素の例
〈*obj*, を〉	13xxx, 11xxx	返事, 名前, 字, 宛先, 小説
〈*subj*, が〉	12xxx(人間活動の主体)	彼女, 先生, 私
〈*on*, に〉	14xxx	黒板, カード, 紙
〈*with*, で〉	14530(文具)	筆, 鉛筆

に抽出できる可能性がある[67]. 現在, 和英辞典から取り出した約4万の対訳例を対象にして実験を行っている. 「書く」という動詞が現われた約200の例文の解析結果から取り出された格フレームを表-2に示す.

大規模な語彙についてこのように人手によってもしくは計算機援用によって現在得られるものは, 各単語にとってみるときわめて希薄な知識に過ぎない. 各語については, その構文的情報ないし用法だけでなく, 意味的に詳細な情報も必要である. 自然言語解析用の辞書の最終的な姿は, 現在考えられているどの知識表現システムよりも詳細な意味記述をもったものになるに違いない. 自然言語解析のための知識表現言語の設計もいくつか行われている[1),43),50)]. また, 自然言語処理だけを対象にしたものではないが, 知識表現の標準化という動きがある[46)]. 知識の表現の基本的な問題は, 表現法の定義およびそれを用いた推論機構の定式化である. しかし, それとは別に, 大規模な知識表現システムの一貫性を保った構築をいかに行うかという大きな問題がある. 人間が培ってきた知識を記述するための最も一般的な道具は自然言語である. 頑健な自然言語処理システムの存在は知識表現のボトルネックの解消のための一つの鍵であると言えよう.

5. おわりに

頑健な自然言語処理について著者が重要と考えることについて解説した. システムの能力を拡大して, 多様な言語現象を柔軟に扱わねばならない. たとえ, 構文的または意味的に誤っているとしても人間が理解可能であるかぎり, そして, 人間がそのような誤りを犯すかぎり, 処理システムはそれを理解しなければならない. また, 辞書や知識の表現なども自然言語処理に用いるかぎりはその量の問題に対処しなければならない. コーパスを対象とした言語現象解析には現在はまだ統計的な方法が主として用いられている. 本解説では, 自然言語理解への方向性に主眼をおき, この方向の研究については最小限の記述にとどめた. 今後, 大規模コーパスのより深い解析の研究に期待したい.

さまざまな言語現象に対応した解析法や, 不適格文の解析法についても優れた手法が数多く考えられている. 今後, これらの技術を用いて, いかに見通しよく大規模なシステムを作っていくことができるかが問題である. その際に, システムのモジュール性, 単調性, 漸進性, 学習機能などが重要なポイントになるのではないだろうか.

大規模な文法の開発はもちろんいくつもの組織で行われている(たとえば, 文献22), 25), 54)). しかし, これらを研究者の間で共有化するのは難しい. 一方で, Alvey プロジェクトで作られた文法や自然言語ツール (Alvey Natural Language Tools, ANLT) を有料ではあるが公開しようという動きがある*. われわれも ICOT と共同で, 日本語用の辞書や文法を含めた自然言語ツールを公開しようとしている**. OS やその他のソフトウェアが多くのユーザの使用によって支えられるように, 自然言語データやツールの共有化は頑健な自然言語処理システム実現への必要不可欠の重要事項ではないだろうか.

参考文献

1) Allen, J. F.: The RHET System, *SIGART Bulletin*, Vol. 2, No. 3, pp. 1-7 (1991).
2) Brent, M. R.: Automatic Acquisition of Subcategorization Frames from Untagged Text, In *Proceedings of ACL-91*, pp. 209-214 (1991).
3) Carbonell, J. G. and Hayes, P. J.: Recovery Strategies for Parsing Extragrammatical Language, *Computational Linguistics*, Vol. 9, No. 3-4, pp. 123-146 (1983).
4) Carbonell, J. G. and Hayes, P. J.: Coping with

* 問い合わせ先: Ted Briscoe (briscoe@cl.cam.ac.uk)
** 問い合わせ先: 松本 (matsu@kuee.kyoto-u.ac.jp), 佐野 (sano@icot.or.jp)

Extragrammaticality, In *Proceedings of COLING-84*, pp. 437-443 (1984).
5) Carbonell, J. G. and Hayes, P. J.: Robust Parsing Using Multiple Construction-Specific Strategies, in *Natural Language Parsing Systems*, Leonard Bolc (ed.), pp. 1-32, Springer-Verlag (1987).
6) Church, K. and Hanks, P.: Word Association Norms, Mutual Information and Lexicography, In *Proceedings of ACL-89* (1989).
7) Dagan, I. et al.: Two Languages are More Informative than One, In *Proceedings of ACL-91*, pp. 130-137 (1991).
8) Dahl, V. and McCord, M. C.: Treating Coordination in Logic Grammars, *Americal Journal of Computational Linguistics*, Vol. 9, No. 2, pp. 69-91 (1983).
9) Eastman, C. M. and McLean, D. S.: On the Need for Parsing Ill-formed Input, *American Journal of Computational Linguistics*, Vol. 7, No. 4, p. 257 (1981).
10) Erbach, G.: Lexical Representation of Idioms, IBM Germany Science Center, IWBS Report 169 (1991).
11) Fauconnier, J., 坂原 茂他訳：メンタル・スペース，白水社 (1987).
12) Fass, D. and Wilks, Y.: Preference Semantics, Ill-Formedness, and Metaphor, *Computational Linguistics*, Vol. 9, No. 3-4, pp. 178-187 (1983).
13) Fink, P. K. and Biermann, A. W.: The Correction of Ill-Formed Input using History-Based Expectation with Applications to Speech Understanding, *Computational Linguistics*, Vol. 12, No. 1, pp. 13-36 (1986).
14) Fong, S. and Berwick, R. C.: New Approach to Parsing Conjunctions Using Prolog, In *Proceedings of ACL-85*, pp. 118-126 (July 1985).
15) Gazdar, G.: *Generalized Phrase Structure Grammar: A Theoretical Synopsis*, Indiana University Linguistic Club (Aug. 1982).
16) Gross, M.: Lexicon-Grammar, The Representation of Compound Words, In *Proceedings of COLING-86*, pp. 1-6 (1986).
17) 春野雅彦，松本裕治，長尾 真：痕跡を扱うためのチャート法の拡張，情報処理学会，自然言語処理研究会 89-5 (1992).
18) Hayes, P. J. and Mouradian, G. V.: Flexible Parsing, *American Journal of Computational Linguistics*, Vol. 7, No. 4, October-December, pp. 232-242 (1981).
19) Hendrix, G. G.: Human Engineering for Applied Natural Language Interface System, In *Proceedings of IJCAI-77*, pp. 183-191 (1977).
20) Hindle, D.: Noun Classification from Predicate-Argument Structures, In *Proceedings of ACL-90*, pp. 268-275 (1990).
21) Hirschman, L.: Conjunction in Meta-Restriction Grammar, *Journal of Logic Programming*, Vol. 3, No. 4, pp. 299-328 (1986).
22) Hirschman, L. et al.: The PUNDIT Natural-Language Processing System, *AI Systems in Government Conference*, IEEE (1989).
23) Jensen, K. and Heidorn, G. E.: The Fitted Parsing: 100% Parsing Capability in a Syntactic Grammar of English, In *Proceedings of Conf. on Applied Natural Language Processing*, pp. 93-98 (1983).
24) Jensen, K. et al.: Parse Fitting and Prose Fixing: Getting a Hold on Ill-Formedness, *Computational Linguistics*, Vol. 9, No. 3-4, pp. 147-160 (1983).
25) Jensen, K. and Bonot, J-L.: Disambiguating Prepositional Phrase Attachments by Using On-Line Dictionary Definitions, *Computational Linguistics*, Vol. 13, No. 3-4, pp. 251-260 (1987).
26) 情報処理振興事業協会技術センター：計算機用日本語基本動詞辞書 IPAL (Basic Verbs) 説明書 (1987).
27) 情報処理振興事業協会技術センター：計算機用日本語基本形容詞辞書 IPAL (Basic Adjectives) 説明書 (1990).
28) Kaplan, R. M. and Bresnan, J.: Lexical-Functional Grammar: A Formal System for Grammatical Representation, Chap. 4 of *The Mental Representation of Grammatical Relations*, Bresnan, J. (ed.), MIT Press, pp. 173-281 (1982).
29) 加藤恒昭：非文の解析 —チャートに基づく新たな手法，情報処理学会，自然言語処理研究会 83-10 (1991).
30) Kay, M.: Algorithm Schemata and Data Structures in Syntactic Processing, XEROX PARC, CSL-80-12 (1980).
31) Kay, M.: Functional Unification Grammar: A Formalism for Machine Translation, In *Proceedings of COLING-84*, pp. 75-78 (July 1984).
32) Klavans, J. and Tzoukermann, E.: The BICORD System: Combining Lexical Information from Bilingual Corpora and Machine Readable Dictionaries, In *Proceedings of COLING-90*, pp. 174-179 (1990).
33) 今野 聰，田中穂積：左外置を考慮したボトムアップ構文解析システム，コンピュータソフトウェア，Vol. 3, No. 2, pp. 115-125 (1986).
34) Kurohashi, S. and Nagao, M.: Dynamic Programming Method for Analyzing Conjunctive Structures in Japanese, In *Proceedings of COLING-92* (1992).
35) Kwansy, S. C. and Sondheimer, N. K.: Relaxation Techniques for Parsing Grammatically Ill-Formed Input in Natural Language Understanding Systems, *Computational Linguistics*, Vol. 7, No. 2, pp. 99-109 (1981).
36) Lakoff, M. and Johnson, M.: *Metaphors We Live By*, University of Chicago Press (1980).
37) Lavelli, A. and Stock, O.: When Something is Missing: Ellipsis, Coordination and the Chart,

In *Proceedings of COLING-90*, Vol. 3, pp. 184-189 (1990).
38) Lenat, D. B. et al.: CYC: Toward Programs with Common Sense, *CACM*, Vol. 33, No. 8, pp. 30-49 (1990).
39) Lenat, D. B. and Guha, R. V.: *Building Large Knowledge Bases*, Addison-Wesley, Mass. (1990).
40) Levinson, S. C., 安井 稔他訳: 英語語用論, pp. 118-119, 研究社出版 (1990).
41) Matsumoto, Y., Yamagami, K. and Nagao, M.: Bidirectional Parsing for Idiom Handling, In *Proceedings of IJCAI-91 Workshop on Nonliteral Language*, pp. 83-91 (1991).
42) Mellish, C.: Some Chart-Based Technique for Parsing Ill-formed Input, In *Proceedings of ACL-89*, pp. 102-109 (June 1989).
43) Miller, B. W.: The Rhetorical Knowledge Representation System, Reference Manual (for Rhet Version 17. 9), Technical Report 326, University of Rochester, Computer Science Department (1990).
44) 長尾 確, 丸山 宏: 自然言語処理における曖昧さとその解消, 本特集, pp. 746〜pp. 756 (1992).
45) 長尾 眞他: 科学技術論文における並列句とその解析, 情報処理学会, 自然言語処理研究会 36-4 (1983).
46) Neches, R.: Enabling Technology for Knowledge Sharing, *AI Magazine*, Vol. 12, pp. 36-56 (1991).
47) 日本電子化辞書研究所: 概念辞書 (第2版), TR-012 (1989).
48) Pereira, F. C. N. and Warren, D. H. D.: Definite Clause Grammars for Language Analysis—A Survey of the Formalism and a Comparison with Augmented Transition Networks, *Artificial Intelligence*, Vol. 13, pp. 231-278 (1980).
49) Pereira, F.: Extraposition Grammars, *American Journal of Computational Linguistics*, Vol. 7, No. 4, pp. 243-256 (1981).
50) Pletat, U.: The Knowledge Representation Language LILOG, in *Text Understanding in LILOG*, Herzog, O. and Rollinger, C.-R. (eds.), Lecture Notes in Artificial Intelligence 546, Springer-Verlag, pp. 357-379 (1991).
51) Pollard, C. and Sag, I. A.: Imformation-Based Syntax and Semantics, Vol. 1 Fundamentals, CSLI Lecture Notes, No. 13 (1987).
52) Pratt, V. R.: LINGOL—A Progress Report, In *Proceedings of IJCAI-75*, pp. 422-428 (1975).
53) Rau, L. F.: Robust Natural Language Processing, In *Proceedings of PRICAI-90*, pp. 95-100 (1990).
54) Robinson, J. J.: DIAGRAM, *CACM*, Vol. 25, No. 1, pp. 27-46 (1982).
55) Sacerdoti, E. D.: Language Access to Distributed Data with Error Recovery, In *Proceedings of IJCAI-77*, pp. 196, 202 (1977).
56) Sadler, V.: The Textual Knowledge Bank: Design, Construction, Application, In *Proceedings of the 1st International Workshop on Fundamental Research for the Future Generation of Natural Language Processing* (*FGNLP*), pp. 17-32, ATR (1991).
57) Schank, R. C.: *Conceptual Information Processing*, North Holland (1975).
58) Sedogbo, C.: A Meta-Grammar for Handling Coordination in Logic Grammars, in '*Natural Language Understanding and Logic Programming*' Dahl, V. and Saint-Dizier, P. (eds.), North Holland, pp. 65-78 (1985).
59) Small, S.: Viewing Word Expert Parsing as Linguistic Theory, In *Proceedings of IJCAI-81*, pp. 70-76 (1981).
60) Stock, O.: Parsing with Flexibility, Dynamic Strategies, and Idioms in Mind, *Computational Linguistics*, Vol. 15, No. 1, pp. 1-18 (1989).
61) 田中穂積他: 自然言語処理のためのプログラミングシステム—拡張 LINGOL について—, 電子通信学会論文誌, J 60-D, pp. 1061-1068 (1977).
62) Thompson, B. H.: Linguistic Analysis of Natural Language Communication with Computers, In *Proceedings of COLING-80*, pp. 190-201 (1980).
63) 徳永健伸, 岩山 真, 田中穂積: 論理文法におけるギャップの扱い, 情報処理学会論文誌, Vol. 32, No. 11 (1991).
64) 冨浦洋一他: 語義文からの動詞間の上位-下位関係の抽出, 情報処理学会論文誌, Vol. 32, No. 1, pp. 42-49 (1991).
65) 鶴丸弘昭他: 国語辞典情報を用いたシソーラスの作成について, 情報処理学会自然言語処理研究会 83-16 (1991).
66) Utsuro, T., Matsumoto, Y. and Nagao, M.: Lexical Knowledge Acquisition from Bilingual Corpora, In *Proceedings of COLING-92* (1992).
67) 宇津呂武仁, 松本裕治, 長尾 眞: 二言語対訳コーパスからの動詞の格フレーム獲得, 人工知能学会第6回全国大会論文集 (1992).
68) Weischedel, R. M. and Black, J.: Responding Intelligently to Unparsable Inputs, *American Journal of Computational Linguistics*, Vol. 6, No. 2, pp. 97-109 (1980).
69) Weischedel, R. M. and Sondheimer, N. K.: Metarules as a Basis for Processing Ill-Formed Input, *Computational Linguistics*, Vol. 9, No. 3-4, pp. 161-177 (1983).
70) Wilensky, R. and Arens, Y.: PHRAN—A Knowledge-Based Natural Language Understander, In *Proceedings of ACL-80*, pp. 117-121 (1980).
71) Wilensky, R.: A Knowledge-Based Approach to Language Processing: A Progress Report, In *Proceedings of IJCAI-81*, pp. 25-30 (1981).
72) Woods, W. A.: Transition Network Grammars for Natural Language Analysis, *CACM*, Vol. 13,

No. 10, pp. 591-606 (1970).
73) Woods, W. A.: An Experimental Parsing System for Transition Network Grammars, in *Natural Language Processing*, Rustin, R. (ed.), pp. 111-154, Algorithmics Press (1973).
74) Zernik, U. and Dyer, M. G.: The Self-Extending Phrasal Lexicon, *Computational Linguistics*, Vol. 13, No. 3-4, pp. 308-327 (1987).

(平成4年5月26日受付)

松本 裕治 (正会員)

昭和30年生. 昭和52年京都大学工学部情報工学科卒業. 昭和54年同大学院工学研究科修士課程情報工学専攻修了. 同年電子技術総合研究所入所. 昭和59～60年英国インペリアルカレッジ客員研究員. 昭和60～62年(財)新世代コンピュータ技術開発機構に出向. 昭和63年京都大学大型計算機センター助教授. 平成元年京都大学工学部電気工学第2学科助教授となり, 現在に至る. 自然言語処理, 論理プログラミング等に興味を持つ.