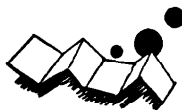


解説



科学技術庁機械翻訳プロジェクトの概要†

長尾 真^{**} 辻井 潤一^{**} 中村 順一^{**}
坂本 義行^{***} 鳥海 剛^{****} 佐藤 雅之^{****}

1. はじめに

言語の機械翻訳は長年の夢であったし、現在もそうである。言語は人間の頭脳活動の最も具体的な現れであり、言語活動は人間の頭脳活動のメカニズムが解明されない限り、十分に解明されるにはならないだろう。したがって、言語の機械翻訳という問題に直接取り組むことは無謀であるといえることができる。しかし1950年代の終りごろから始められた機械翻訳の研究は、途中うねり曲折はあったものの、多くのことを明らかにしてきた。したがって、これを総合化してしっかりしたシステムを作れば、ある程度使えるシステムが作れるだろうという期待をここ十年間くらい前から持っていたのである¹⁾。しかしそれを具体的に示すためには、いくつかの条件が必要であった。それは、強力な計算機処理能力をさせること、かなりの人数の研究者による数年間の集中した研究開発作業が必要なこと、そのための研究費と研究体制をどうするかという問題であり、これは1大学の1研究室ではできないものではなかった。

一方、機械翻訳に対する社会的要請は、ここ数年非常に大きくなってきており、また日本の科学技術活動の実態を諸外国に知らせることが、国としても必要となってきた。幸いにも国はこのような状況をよく認識し、科学技術会議は科学技術振興調整費の1部を用いて昭和57会計年度から3年間(その後予算などの都合によって4年間に修正された)にわたって、「日英科学技術文献の速報システムに関する研究」を発足

させることを決定した。この研究は諸外国との科学技術文献交流促進の必要性に鑑み、それら文献(抄録)の翻訳を効率的に行うため、

(1) 科学技術用語辞典データベースを用いて科学技術文献を翻訳するための構文パターン変換方式による日英(英日を含む、以下同じ)言語処理ソフトウェアの開発。

(2) 科学技術分野の専門用語を収集、整理し、コンピュータで利用可能な形に編集した日英科学技術用語辞書データベースの開発。

(3) これらを統合して利用しうる翻訳システムの開発を行う。

を目的としたものである。

この目的にそって、科学技術文献の日英翻訳システム、英日翻訳システムの2つのシステムを作ることになった²⁾。文献の分野としては種々の検討の末、われわれシステム開発者が理解しやすい電気工学分野をとることにし、日本語文献としては日本科学技術情報センターの電気工学分野の文献速報のタイトルおよび抄録文、英語文献としてはINSPECの同じ分野の抄録をとることにした。

研究開発グループとしては、京都大学、電子技術総合研究所(ETL)、日本科学技術情報センター(JICST)、工業技術院筑波情報計算センター(RIPS)の4者が協力して研究開発することになった。研究開発の分担は次のようである。京都大学は機械翻訳用ソフトウェアの開発、解析・変換・生成の文法規則の開発、解析・変換・生成に用いる辞書情報の種類の決定、各単語に対してそれらの情報を与える作業基準の作成、変換部分の辞書データの1部の作成などを担当した。ETLは形態素解析、形態素生成のソフトウェアとその辞書の作成、解析のための用言辞書の作成を担当した。JICSTは主として体言辞書の作成を担当した。体言は一般用語のほかには大量の電気工学分野の専門用語を含む。RIPSは京都大学が作る機械翻訳システムにテキ

† Outline of Machine Translation Project of the Science and Technology Agency by Makoto NAGAO, Jun-ichi TSUJII, Jun-ichi NAKAMURA (Dept. of Electrical Engineering, Kyoto University), Yoshiyuki SAKAMOTO (Electrotechnical Laboratory), Tsuyoshi TORIUMI, Masayuki SATO (The Japan Information Center for Science and Technology).

** 京都大学工学部電気工学第二教室

*** 電子技術総合研究所

**** 日本科学技術情報センター

スト入力部分、テキスト出力部分、対話的修正部分、辞書の保守に関するソフトウェアなどを付け加え、機械翻訳のための総合システムを作ることになっている²⁾⁻⁹⁾。

3年半あまり経過し、現在日英翻訳システムはほぼ出来上がり、翻訳の質の評価をしつつある。英日翻訳システムは1984年度から本格的に作成しはじめ、現在ある程度の翻訳結果を出せるところまできたが、今後残された期間にできるだけの改良を加える予定になっている。したがって、本論文ではほぼ完成している日英翻訳システムについて、その概要を述べることにする。

2. システム設計の考え方

かなり複雑な文に対しても対処できるような機械翻訳システムを非常な短期間に4つの研究機関の協力によって作らねばならないということは非常に難問であった。現状の技術の統合によってシステムを実現することが目標で、それ以上の研究的要素はなるべく持ち込まないということが前提とされたが、小さなモデルで少数の文を翻訳する研究のレベルとはまったく異なったシステム開発であるから、未知の部分はかなりあり、それらに対しても対処できるシステムとしなければならないなどのことからシステム設計の考え方として次のような方針をとった²⁾。

(1) 未知語や未知の構造の存在によって処理を途中で放棄することなく、不完全であっても、最後まで処理を行って翻訳結果を出すシステムであること。これは翻訳文の後修正をする場合の大切な考え方である。

(2) 研究開発の発展途上において明らかになってくるであろう未知の言語現象にも十分対処できるように、機械翻訳用のソフトウェア・システムは強力な表現力を持つように設計する。ソフトウェアは解析・変換・生成のすべての段階で用いることができ、種々の異なった性質の言語が扱えるように処理過程の制御がきめこまかくできるシステムを作る。

(3) 解析・変換・生成、それぞれが膨大な規則になるために、多人数で並行に開発ができる必要がある。したがって、部分文法という考え方を導入し、部分文法ごとに独立に開発できるようにする。こうすることによって、翻訳の誤り原因の発見なども早くなり、システム全体の管理がしやすくなる。

(4) 将来の多言語間機械翻訳の可能性を考えて、

解析と変換、生成の3つの段階を明確に分ける。それらの段階をつなぐインタフェースの構造をできるだけ明らかにする。

(5) 言語は少数の基本的な規則によって説明されるものではなく、膨大な例外によって成り立っているものである。したがって、できるだけ広い言語現象に対して適用できる規則を探すことは当然であるが、各単語に固有の言語現象がうまく扱えるようなメカニズムを設けることが必要となる。われわれの場合は、辞書の中に各単語特有の文法規則を書けるようにし、それを処理過程において優先的に使えるメカニズムを作ること考えた。

(6) 辞書情報としては、現在の機械翻訳に直接必要なものだけでなく、将来必要となるであろう情報をも入れた、できるだけ中立的で理想的な形のものを作ることが必要である。一方、大量の単語に対して正確な情報を与えるためには、複雑な辞書構造をとることはよくない。さらに単語固有の言語現象を十分に記述できることも必要である。このような矛盾する要求に対処する方策を考えた。

(7) 言語の解析の基本には格文法の考え方をとる。これは日本語の解析に特に適しているだけでなく、英語その他のヨーロッパ諸言語の解析にも広く用いられてきつつある方法である。格文法をとるために格要素の設定と、各格要素にどのような名詞がはいりうるかを記述するための意味素性の集合を決定しなければならない。これは日本語のほか英語、その他のヨーロッパ諸言語の特性を知って作る必要がある。

3. ソフトウェアシステム

前章で述べた種々の言語の記述に関する要請から、文法記述のためのソフトウェアシステムとしては、木構造を木構造に変換する規則を採用することにした。木構造の各ノードは任意の数の属性・属性値対をもつことができるようになっている。この部分に性・数・格の一致、意味の整合性などを記述する。このように複雑な木構造を取り扱い、しかも文法規則、辞書構造とともに定形的なものにはできないということから、基本となるプログラム言語としてLISPを採用することにした。LISPの採用については、PL/1, PASCALなどの他の言語との比較検討をしたが、今回の研究においては翻訳速度よりは言語現象を十分に記述することに重点を置いたこと、さらに短期間にこのソフトウェアシステムを開発しなければならないことなどの

要因から LISP の採用を決定した。

われわれの開発した GRADE (Grammar Descriptor)¹⁰⁾と名づけた文法記述言語の書き換え規則の部分は次のような構造を持っている。文法を表現する書き換え規則としては、「左辺の木→右辺の木」という形に書かれる。この規則の左辺の木は、翻訳のあらゆる中間段階にある翻訳対象文の構造（この節ではこれを単にデータと呼ぶ。これも一般には木構造をしている）との間でマッチングがとれるかどうか調べられ、データのある部分でマッチングがとれたら、その部分が規則の右辺の木構造に書き換えられる。

木構造のマッチングにおいて次のような機能が存在する。(1)属性・属性値対がついていれば、それもマッチングの対象となる。(2)変数を木構造のノードに書くことができる。(3)木の枝の左右の順序が任意であることを指定できる。(4)規則の左辺の木のマッチングに関して複雑な条件が必要な場合には、それを LISP プログラムで書くことができる。(5)この規則の部分で単語辞書内に書かれている単語固有の規則を呼び出し動作させることができる。

文法規則の数は、たとえば日本語の解析文法を考えたとしても数 100 から 1000 近くになるから、解析のあらゆる中間段階でこれらすべての規則の適用可能性を調べるということは大変な手間になる。また、名詞句を解析するための規則、文の構造を決定するための規則などは自ずから適用される時点が違っていて、言語学的に自然な順序で解析を進めるのがよい。したがって、文法規則すべてを一様に扱うのではなく、異なった文法現象に関する規則をそれぞれ別々に集めて、これらを部分文法と名付けることにした。単純名詞句を解析する部分文法、並列名詞句を解析する部分文法、埋め込み文構造を決定する部分文法などである。そして、1つの部分文法を1つのノードに対応させ、ノードを有向グラフ構造につないで部分文法の適用順序を規定することにした。これを部分文法ネットワークと呼んでいる。上記の名詞句、並列句、埋め込み文などの処理部分は、実際はすべてこの部分文法ネットワークとなっている。文法全体は部分文法の組み合わせから構成される。

そこで、GRADE にはいくつかの書き換え規則をまとめて1つの部分文法とする機能、その部分文法の実行後、次に適用すべき部分文法を条件つき指定する機能をもっている。さらに1つの部分文法の中の書き換え規則の適用の仕方を指定することもできる。すなわ

ち、書き換え規則をデータの左から適用してゆくか、右から適用してゆくかの指定、さらにマッチングが成功して書き換えが行われたあと、さらに同一規則の適用の可能性を調べるか、あるいは成功した規則を用いず、次の規則を試みるかという違いを指定できる。書き換え規則の表現の仕方と規則適用のモードとの間には関係があって、うまくこれらを調整すれば能率のよい解析ができるのである。

文の解析を行うと通常非常に多くのあいまいさが生じる。すなわち、1つのデータに異なったいくつかの構造を対応させることになる。われわれの場合には、できる限り広い範囲を一度にみて意味のあいまいさの生じることを防いでいるが、それでもあいまいな構造は生じる。文法規則を書くときに、ある構造と別のある構造が同じ部分に生じた場合に、それらは同じ句として解釈されることを積極的に表現することができるように PARA ノードと称する特別なノードを作り、システム側がその解釈を自動的に行うよう GRADE を作ってある。

GRADE の制御部は2つの部分からなる。1つは人間にわかりやすい形で書かれた上記の書き換え規則をパターンマッチングしやすい内部表現に変換する変換部であり、もう1つは内部表現に変換された書き換え規則をデータに適用して処理を進める実行部である。これらはすべて LISP で書かれていて種々の支援プログラムも入れて1万行をはるかに越える複雑なプログラムとなっている。

GRADE には、この他に辞書システムとの結合をする部分がある。辞書データは膨大であるので現在のところ大型計算機の標準ランダムアクセス・ファイルシステムを利用して記憶されている。したがって、必要な辞書データはファイルから読み出して LISP 形式で主記憶の上に展開し GRADE 文法の使用に供するようになっている。

GRADE に付属した支援プログラムには種々のものがある。辞書データの誤りをできるだけ少なくするために、パンチされた辞書データの誤りをデータの初期入力の際にできるだけ発見するプログラムが作られており、また、計算機内の辞書データの変更、追加などのためのプログラムなどが代表的なものである。4つの研究機関が文法、辞書データ作成を分担しているため、計算機間のコード変換をはじめとして、お互いのインタフェースをうまくとるためのプログラムがいろいろ必要で、こういったなんでもないとこに予

想以上の手間がかかった。GRADEの詳細は他にゆずる^{4),10)}。

4. 辞書情報とそのシステム化

機械翻訳において辞書が大切なものであることはますます認識されてきつつある。われわれの場合には、短期間の間に大量の辞書データを整備蓄積しなければならなかったために、いくつかの工夫を必要とした。それは、辞書データをできるだけ目でわかる形の形式に表し、むだな情報の記入をせず、最小限度の記入ですむようにすること、計算機のことを知らない一般の人でも記入ができることであり、さらに多くの人に並行に作業をしてもらうときにできるだけ質をそろえるために、辞書情報の記入のための作業手続きを詳しく書いたガイドラインを作成したことにある。

辞書は使用の上から入力文解析のための辞書、言語間変換のための辞書、出力文生成のための辞書、およ

び形態素解析のための辞書、形態素生成のための辞書の5つに分けられる。これらの辞書のうち、言語間変換の辞書の部分を除いて、他は1つの中立的な形の辞書として作ることができ、そこから使用目的に応じた辞書を生成して使用する。したがって、単一言語の辞書としては、この中立的な辞書を作ることになり、辞書の管理、すなわち辞書データの追加、変更、削除などはすべてこの中立的な辞書において行うことにしている。

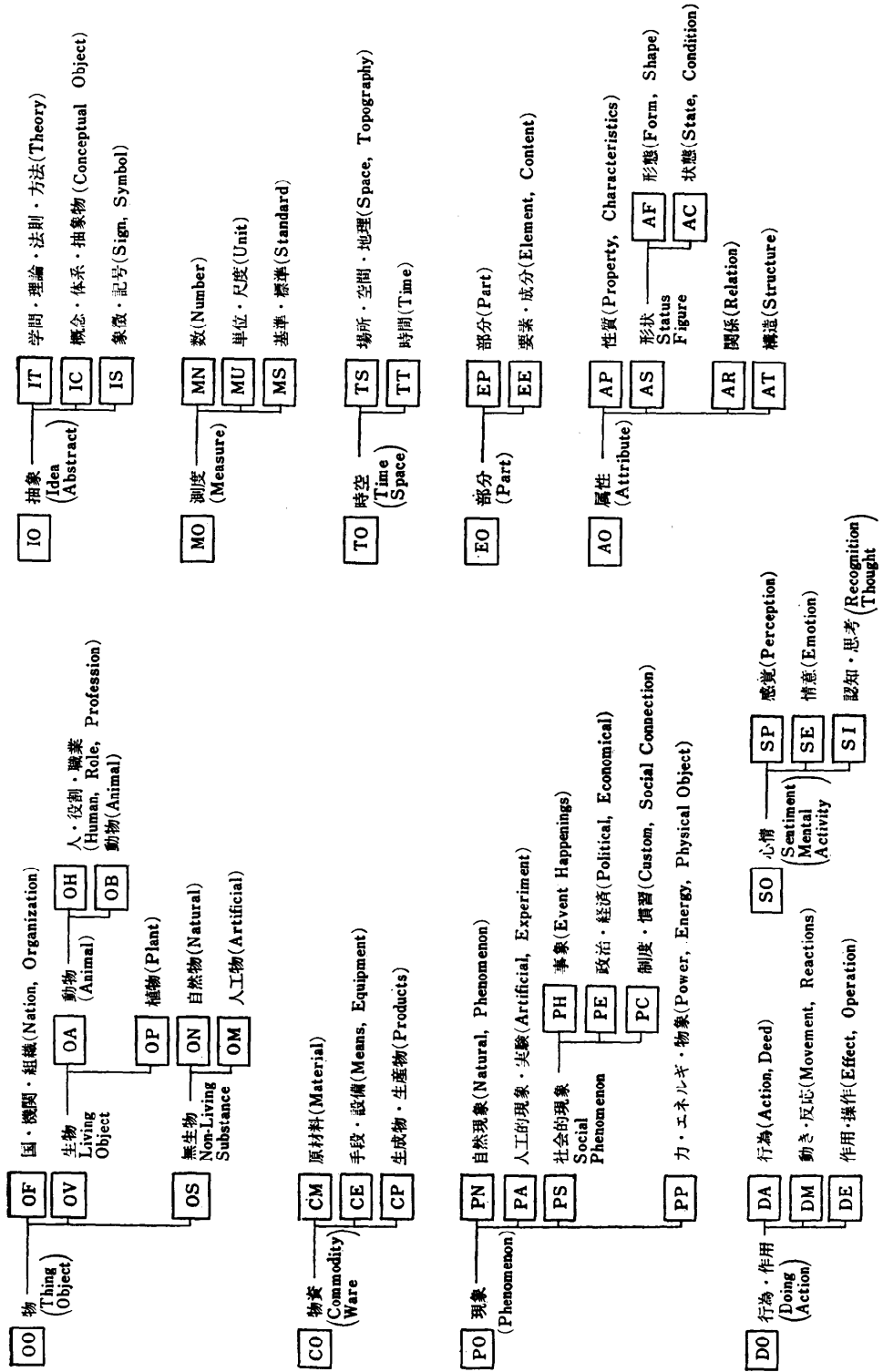
辞書データは現在約55,000語計算機に入れてあるが、その内容は動詞約3,000語、形容詞、形容動詞約1,000語、普通名詞および電気工学を主とする科学技術用語約5万語、連体詞、副詞その他の付属語約800といったところである。動詞、名詞の主なもの翻訳の対象としたJICSTの抄録の最初の3,000文に現れるものを取り、それにJICSTで別途抽出した11万語の用語辞書から選んで作成した。

表-1 日本語格ラベル一覧表

日本語名	英語名	用 例
(1) 主体	SUBject	～が
(2) 対象	OBject	～を
(3) 受け手	RECipient	～に与える
(4) 与え手	ORigin	～から受ける, 奪う
(5) 相手1	PARtner	～と協議する, 異なる, ～に関連する
(6) 相手2	OPPonent	～から保護する, 独立する
(7) 時	TIME	1980年に
(8) 時・始点	Time-FRom	5月から
(9) 時・終点	Time-TO	来年まで
(10) 時間	DURation	5分間加熱する
(11) 場所	SPAce	～に位置する, ～で発生する
(12) 場所・始点	Space-FRom	～から帰る
(13) 場所・終点	Space-TO	～へ送る, ～に到達する
(14) 場所・経過	Space-THrough	～を通る, 上空を飛ぶ
(15) 始状態	SOUrce	↓ 5.5%から6%へ引き上げる
(16) 終状態	GOAL	↑ 英語から日本語に翻訳する
(17) 属性	ATTRIBUTE	適応性に富む, 欠ける, 乏しい
(18) 原因・理由	CAUse	事故で死ぬ, ～から分かる
(19) 手段・道具	TOOL	イオン法で, ドリルで
(20) 材料	MAterial	ペーストで作る
(21) 構成要素	COMponent	～から成る, ～で構成する
(22) 方式	MANner	並列に, 10 m/secで
(23) 条件	CONdition	焦点深度で決まる
(24) 目的	PURpose	～に適する, 備える, 必要な
(25) 役割	ROLe	議長に選ぶ, ～として用いる
(26) 内容規定	COntent	～と呼ぶ, 述べる, みなす
(27) 範囲規定	RANge	～について, ～に関して
(28) 提題	TOPic	～は, ～とは
(29) 観点	VIEWpoint	立場から, ～の点で
(30) 比較	COmparison	～より大きい, ～に劣る, ～を上回る
(31) 随伴	ACOMpany	～とともに, ～に伴って
(32) 度合	DEGREE	5%増加する, 3キロやせる
(33) 陳述	PREdicative	～である

注) 英語名中, 大文字の部分(3字)を略称とする。

表-2 名詞意味マーカー一覧表



辞書に記入すべき情報はどのような文法概念を用いてどのような解析を行うかということに依存することはやむをえないが、できるだけ中立的、一般的に作ることを心がけた。文法としては格文法をとることにし、格の種類として表-1に示すものを採用した。ある動詞の種々の格スロットにどのような名詞が入りうるかを規定する方法としては、意味素性の組み合わせによる方法を取り、そのための意味素性としては表-2に示すものをとった。これらの意味素性のいくつかのもの AND, OR 結合によって、動詞の格スロットに入りうる名詞を規定する。

辞書情報は記入フォーマットをきめ、記入のための作業基準書を作って多人数での作業が可能なよう工夫をした。記入フォーマットは名詞、動詞、形容詞、副詞その他の品詞について作られている。辞書は日本語(解析, 生成に共用)辞書, 日英変換のための辞書, 英語(解析, 生成)辞書などが用意されているが, その詳細は他にゆずる^{7), 8), 11)}。

5. 日本語文の解析

(i) 形態素解析³⁾

日本語文の解析は非常に面倒である。形態素解析は構文解析と切りはなされていて、LISP プログラムでその処理内容が書かれている。まず、単語の切れ目として可能な部分で文字列を分割し、辞書を引き、品詞活用形などを決める。そして隣接する品詞(と活用など)の接続可能性を接続表によってチェックし、許されていればその分割は正しいとし、許されていないならばその分割は正しくなかったとして分割の位置を変更する。複数の異なった分割が許され、それらは 3. で述べた PARA ノードによって保持される。

文の分割が困難なのは、予想しない文字列が現れるからである。文の中に数式が現れたり、化学式が挿入されていたりして、種々の記号が異なった意味で使用されるので接続表を完全なものにするのは非常に困難である。分割のプログラムと接続表は分離されているので、接続表だけの改良ですむようになっている。

(ii) 日本語構文解析^{5), 12)}

(a) 日本語の構文解析は図-1に示す部分文法ネットワークの組み合わせに従って行っている。この図の各部分はさらに多くの部分文法に分割されている。このような分割を前提にして各部分の文法規則を作るが、こうすることによって複数人による文法規則の並行開発が可能となった。図中の(A)は構文解析を簡潔

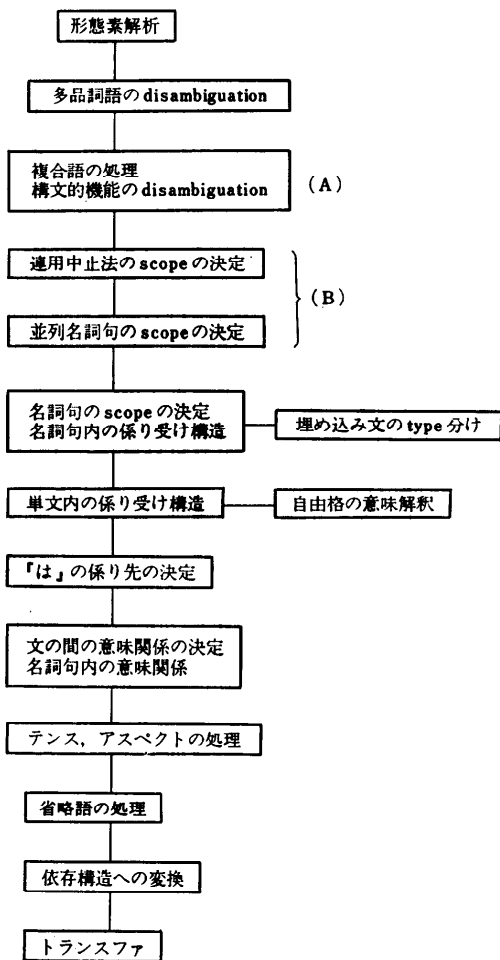


図-1 日本語解析過程の処理例

なものとするための処理で次のような仕事をする。

(1) 1語にまとめた方が後の解析が簡単化されるような隣接語をまとめて連語とする。「数」+「単位」, 「形容動詞語幹」+「名詞」(高速処理など), 格助詞, 副助詞の連続したものなどがその例である。

(2) 周囲の環境を参照して、語レベルのあいまいさを解消したり、品詞の変更をしたりする。たとえば、「で」は格助詞か助動詞「だ」の連用形かを区別することが難しいが、この段階で明確化できる場合には一意の解釈に決定する。形容詞、形容動詞の連用形は副詞として使われることが多いので、句点の存在や、後続の文の形などによって連用中止法の可能性が強いものなどを除いて、副詞として扱う。

(3) 各種の省略の復元

抄録文特有の表現として、動作性名詞は後に続く「する」を省略することが多いので、これを復元する。また、「～が〈形容動詞語幹〉の場合」などでは「の」のかわりに「である」が存在するものと変更する。このような類のものはほかにも多い。

(b) Bの文や句の並列の処理では、どの範囲の語句が並列するかを決定する必要がある。従来多くの解析では、これは可能なあらゆる解析という部分に押しやっていた。しかし並列語句には多くの表面上のキーが存在する。たとえば、同じ単語や語基が存在すること、関連性の高い語が存在すること、「～と～と」、「～、～など」、「～と～の一致」などの並列を示す表現があること、運用中止法の範囲の決定には、動詞のもつ意志性、準意志性、無意志性などの特徴が共通に存在することなどである。したがって、構文解析をする前に、これらの文中に現れる表面上の特徴を調べて、並列する語句の範囲を推定し、そのあとそれぞれの部分に対して解析を行う。

この並列性を検出する GRADE の文法規則では、解析を行うというよりは、まず範囲を設定するというように規則が使われる。並列性の決定は意味情報にたよることもできるが、すぐ意味のレベルにゆかず、文の表面上の各種の形態論的、構文論的特徴から通常のほとんどの並列句は解析が可能である。抄録文には種々の並列句が非常に多く現れるので、この部分の解析が適切に行われることは非常に重要であり、現在、100余りの GRADE の規則が並列句の解析のために設けられている。

(c) (b)で並列句の範囲が推定できるので、それぞれの範囲内において解析を行う。1つの範囲の右端には主節の述部が存在するか、名詞句の主名詞が存在する。名詞句の場合に主名詞の左に述部が存在すれば、それは埋め込み文である。いずれの場合にも検出した範囲を左から右に走査し、述語をまずとらえる。その述語が埋め込み文の述語であれば、埋め込み文処理の部分文法にゆく。その述語が主節の述語であれば、単文処理の部分文法にゆく。述語がなければ単純名詞句の処理にゆく。埋め込み文の処理においては述語の右隣りの名詞と述語から左の名詞句とをもって単文処理の部分にゆく。単文処理の部分文法では、まず単純名詞句のまとめを行った後、述語の格フレームとの対応関係を調べて文の構造(格スロット)を決定する。埋め込み文の処理結果はこれらの情報をになった名詞句という形で得られることになる。

単文の処理は述語の左の名詞句を連用形助詞によって区切り、区切られたそれぞれの範囲を名詞句処理の部分文法にわたす。すべての名詞句がこのようにして処理されたあと格関係の決定が辞書内の述語の格情報との関係で決定される。名詞句の構造の決定において「の」を伴う名詞句は右に隣接する名詞句を修飾するものとし、また動作性名詞は複数個の名詞句によって修飾されうる、などの情報によってその構造を決定する。

このように、この(c)の部分は多くの部分文法の再帰的呼び出しによって解析の行われる部分である。多重解析の可能性のある部分については、最も可能性の高い解析結果を優先的に出すことにしている。

(d) 埋め込み文は次の4つのタイプに分類した。

(i) 被修飾名詞が埋め込み文のある格に対応するもの(例: 翻訳した文)

(ii) 被修飾名詞が埋め込み文のある格を埋めている名詞の修飾語句となっているもの(例: 処理速度が速い計算機)

(iii) 埋め込み文が被修飾名詞の具体的内容を示す同格節になっているもの(間に「という」が挿入できる)。 (例: 教師が生徒を殴った事件)

(iv) その他(例: 新しい計算法を理論値と比較した結果、魚を焼くけむり)

これらの各タイプについて、次のような発見的な手がかりを使って文の解釈を行っている。

[タイプ1]

1. 被修飾名詞が表層上「が」・「を」・「に」でマークされる格を埋めることが多い。

2. 述部が受動態の場合には、表層上「が」でマークされる格(深層の目的格)を埋めることが多い。

3. それ以外の格を埋める場合は、被修飾名詞自体が格の名称を示すような特殊な名詞(原因、目的、動機、条件、方法などを含む語)であることが多い。

(例: この装置が正常に動作する環境条件)

4. 3.の特殊な名詞の場合には、「～のための方法」といった補助的表現が挿入されることがある。

[タイプ2]

このタイプは、原理的にはさまざまな形を探りうるが、最も多いのは、

〈部分, 属性, 動作〉が〈状態性述部〉[被修飾名詞]の形式である。これ以外の場合に、このタイプの解釈を許すことは、あいまいさを増やすだけで、全体としての正答率を悪くする(ただし、〈部分, 属性, 動作〉

は、この種の意味マーカを持つ名詞を示す。)

【タイプ3】

1. 「～という」・「～などの」などの補助的表現が埋め込み文と被修飾名詞の間に挿入される。

2. このタイプを採りやすい一連の被修飾名詞がある(例: こと, 場合, 点, 必要性, 状況, 事実, 方法, など)

【タイプ4】

1. このタイプもタイプ2と同様にさまざまなものがあり得るが、良く生起するパターンのみに限定した方がよい。たとえば、次のようなパターンがある。

「～を用いた〈方法, 行為, 生産物〉」

「～を考慮した〈方法, 行為, 測定〉」

2. このタイプを採りやすい一連の被修飾名詞がある。(例: 音, 匂い, 結果, など)

このような発見的な手がかりは、常に正しい結果を与えるとは限らないし、また、手がかりにも正答率の良いものから、低いものまで種々のものがある。また、

「温度を上げる方法は、うまく行かない」(タイプ3)

「温度を上げる一つの方法は、～である」(タイプ1)

のように、発見の手がかりが異なった解釈を与えることもある。したがって、名詞句並列のスコープ決定の場合と同様に、なるべく多くの場合に正しい解釈が出るように、制御の効いた規則適用を行っている。

(e) 以上では、語の係り受け関係の解析を中心に述べてきたが、これ以外に語や文の意味関係、テンス・アスペクト、省略の処理などが行われる。

(i) 自由格: 動詞格フレーム中の格要素については、各動詞の辞書に深層格で解釈が記述されている。しかし、場所格・時間格・随伴格・様態格などの自由格については、各動詞の辞書には記述できないので、それぞれの格助詞ごとに、名詞意味マーカ・述部の素性(アスペクト、意志性など)を参照して深層格の解釈を絞る¹¹⁾。

(ii) 文と文との意味関係: 連用中止、接続助詞などの表わす文の意味関係を、22の深層格関係に解釈する。比較的意味の明確な接続助詞も、「のに」には逆接、目的の2つの場合があり、「ように」には目的、様態、「ながら」には時間的同時、逆接、「ほど」には度合、比較の意味がある、というように、意味的には多義であり、それに応じて英語での訳出形が変化する。この解釈には、接続される2文の述部のアスペクトと意志性の有無が有効な手がかりとして使われている¹¹⁾。

(iii) テンス・アスペクトの処理: 動詞を状態・準状態・継続・瞬間・推移の5つに分類し、この分類とアスペクト形式素の組み合わせで述部全体のアスペクトを決定している。

(iv) 省略の処理: 格要素の省略、「は」の係り先の決定などを経験的な方法を導入して行っている。

6. 日英変換と英語生成過程

6.1 変換過程の概要⁹⁾

日本語入力文の解析結果として依存構造が得られているので、これを出発点として対応する英語の依存構造を計算するのが、変換過程である。変換過程では、次の2つの仕事を行う。

(1) 日本語の各単語がどのような統語的・意味的環境で使われているかを検査することにより、適切な英語での訳語を選択する。

(2) 依存構造は、原言語の文の意味構造を表現するが、原言語の統語的特質も反映しており、日本語と英語のように語族が離れている言語対においては、この構造自体を変換しないと自然な訳文が得られないことも多い。このような言語対に依存した構造の変換もこの変換過程で行う。

変換過程は次の3段階に分かれる。

(i) Pre-Transfer-Loop: 解析で使われた属性のうちで、以降の処理で不要なものを消去したり、日英変換の処理を容易にするための属性の付与を行ったりするほか、日本語独特の構造で、英語にはそれに直接対応する構造がないものを、より中立的な構造に変換する。たとえば、「Xが多い」は「多くのXが存在する」という形に変換し、「～に注意して～する」のように、日本語では述部として独立し従属節を作っているが、英語では副詞表現に縮退する構造や、そのまま英語にすると冗長となる構造、英語では助動詞相当の表現に変換する必要がある構造、などが処理され、英語の依存構造に近い形に変換される。

(ii) 日英変換: 依存構造木の節点について、訳語選択とそれに伴う局所的な構造の変換を行う。依存構造の頂点節点から出発して、その下位節点の変換を木構造に従って再帰的に起動してゆく。

(iii) Post-Transfer-Loop: 英語らしい表現を作り出すために、英語の依存構造を多少手直しする。たとえば、「Aを回転させる」では、使役態が使われ、文字通り変換すると'make A rotate'となるが、選択された英語動詞(rotate)が自他両用の場合には、使役

を外して直接他動詞の表現にする。

以上のうち(i)と(ii)は、こねた訳文を出すために、各種の準慣用的な文体を処理する段階である。ただ、この両段階はかなり個別的な現象を取り扱うものであり、必ずしも必須ではないため、現時点ではあまり活用されていない。今後、より多くの実験を繰り返すことにより充実させてゆく必要がある。変換部分の詳細については他にゆずる^{12), 13)}。

6.2 英語生成過程

英語生成過程は、変換過程から出力された英語依存構造を句構造表現に作り直すとともに、複数/単数の情報の付与、tense/aspect/modal 表現の助動詞による具体化、主語と述語の数の一致、時制の一致、などの処理を行い、文法的に正しい英語文を作る。また、文体面を考慮して、It-that 構文や There 構文などを使用することにより、より理解の容易な英語文を生成する^{12), 13)}。この過程は次の3段階に分けられる。

(1) Pre-Generation 段階

依存構造の段階で行った方が見通しの良い構造変換を行う。たとえば、日本語に多い自動詞構文「Aによって、BがCになった」を英語的な他動詞構文「A makes B C」に訳出するために、原因格や道具格の位置にある句の深層格解釈を英語特有の格 Causal-Actant に付け替える。ここで、「なる」・「make」は自動詞・他動詞の代表であり、実際には、より具体的な動詞が入る。

例：「データ処理の進歩で自動化船が増加した」

--->Advances in data processing increased the number of automated ships.

ただし、この Pre-Generation の段階では深層格の解釈だけが変更され、実際に統語上の主語の位置に移動するなどの構造的な変換は、つぎの Generation 段階で行われる。この種の変換の延長には、「～すれば、使用者は～することができる」を「allows the user to do～」と変換する場合のように、両言語で大きく構造を変える場合があるが、この種の規則の適用条件がまだ未調査であること、かなり分野依存の規則（たとえば、マニュアル）で抄録文には出現しないこと、などの理由から現時点ではほとんど組み込まれていない。また、この Pre-Generation 段階の処理と前述の Post-Transfer-Loop での処理は、いずれも同じ依存構造を対象にした規則であり、どちらの段階で適用するかに任意性がある。今後、日英以外の言語対に拡張することにより、最適な規則の配置を決定してゆく必

要がある。

(2) Generation 段階

依存構造から句構造への変換を行う。依存構造の頂点節点から下位の節点に向かって再帰的に変換の部分文法を適用してゆく。この段階は、ほとんど直接的な処理である。次の処理が行われる。

- a. delimiter の生成
- b. 生成辞書引き
- c. 必須格の生成と句構造での位置の決定
- d. 自由格の生成と句構造での位置の決定
- e. delimiter の消去と S-node の生成
- f. 受身変形、あるいは、代動詞の使用の決定
- g. 動詞句の処理

ここで、f.の処理は、日本語での省略に対処するためのものである。すなわち、日本語で主語が省略されており、解析段階でそれが補われていない場合、英語で受動態を選択する。また、主語も目的語も省略されている場合には、述部の名詞派生形を主語として、代動詞を述部とする構造に変形する。

例：「ヘリウムで調べ次のことがわかった」

--->The examination is made by the helium and the following fact is found.

(3) Post-Generation 段階

これまでの処理で一応英語の構文が生成されているが、このままでは、非常に読み難い英語になっていることがあるため、変形規則を適用してより自然な文体に変換する。主な規則を次に示す。

- a. It-that 変形：主語の位置の補文が長い場合（これをチェックする機能が GRADE にある）に、it-that 構文を使用することによって、読みやすい英文にする。
- b. 並列名詞句、動詞の等位接続の縮約：重文における重複する格要素や、並列名詞句における重複する修飾句を削除したり、代名詞化する。
- c. 関係節の外置変形：主語を修飾する関係節は主語を長くし、読み難い英文を作る。このため、関係節の外置変形を行う。

例：The parameter range in which～is determined.
--->The parameter range is determined in which～.

- d. 関係節の縮約：完全な関係節の中で、含まれる格要素の数が少なく、縮約しても判りにくくならないものを、名詞修飾の分詞句 (ing, ed) にする。

例：The result which was obtained by～

--->The result obtained by～

e. There 構文の使用: exist を使った存在文をより自然な there 構文に変える.

f. 副詞, 前置詞の位置の決定: (2)の Generation の段階で前置詞句や副詞の標準的な位置は決定されているが, 関係節で修飾されているような長い前置詞句を後置したり, あるいは, 接続詞的副詞 (however, などの conjunct) や名詞修飾の副詞 (even, only, etc.) の位置を最終的に決定する.

g. 比較級, 時制などの調整: この段階になると, finite 形で出力すべき述部が決まっているので, 重文, 複文の時制の一致をとる.

h. 数, 冠詞, 人称の決定: まだ決まっていない冠詞や数を決める. ここでは, 前置詞句や関係節で修飾されている名詞の場合は, the をつけるなどの発見的な基準を使う.

i. インタフェースのための変換: 形態素出力のために木構造を線形化し, 英語の屈折 (語尾) 処理に関する情報を付与する.

j. 形態素生成: この部分は形態素生成辞書を参照して行う簡単な部分である.

7. おわりに

この研究は昭和 57 年度から開始され, 現在 3 年余を経て, 日英翻訳システムはほぼでき上がっており, 現在訳文の評価とシステムの改良を行っている¹⁴⁾. 訳文の質の評価については, 本誌別項¹⁵⁾ (翻訳文の質の評価と言語の制限) で述べるので, ここでは省略するが, 翻訳文の質が原文の質に直接的に影響されることは当然である. われわれの場合, 翻訳の対象となる科学技術抄録文が非常に複雑な文体をしていて, 翻訳を特に困難にしている. したがって, 解析, 変換, 生成の各過程において, 膨大な数の文法規則を用意することが必要となった. (解析に約 1000 規則, 変換生成にも約 1000 規則) 翻訳文の例を図-2 に示す. 翻訳上の工夫と訳文の例は文献 13) を参照していただきたい.

この研究は科学技術文献 (抄録) の実用的翻訳を目指した研究であり, 特に複雑な言語現象に対して, どの程度まで翻訳が可能であるかを示すことに力点をおいて開発を進めた. この目的はかなりの程度達成されたと考えているが, その代り, ソフトウェアシステムが少し複雑になりすぎ, 翻訳速度が非常におそくなった. 素直に翻訳できる文章の場合には, かなり長い文でも 10 秒程度の CPU 時間で結果が得られるが, 複雑な文で翻訳が失敗したり, 不完全な翻訳しかできな

▶ この方法は R. Bellman の理論から導いたもので, 解析幾何理論の応用である.

This method is one derived from theories of R. Bellman, and this method is an application of geometric diffraction theories.

▶ 本論は, エネルギー管理制御システムについての現状を報告することが目的で, New England 電話会社の例を挙げている.

As for this theory, aims are to report current conditions about energy management and control systems, and this theory has given examples of New England Telephone Company.

▶ この回路はわずかの修正で, 他の計算機にも接続できる.

This circuit can also be connected with other computers by slight correction.

▶ 計算機で基板のパターン設計を行い, 部品の手配から製品までの各工程を制御するシステムを検討.

The systems are studied which carry out the pattern design of substrates by computers and control each process from arrangements of parts up to products.

▶ 計算機で基板のパターン設計を行い, 部品の手配から製品までの各工程に与える効果を検討.

The pattern design of substrates is carried out by computers, and effects on each process from arrangements of parts up to products are studied.

▶ 電力用半導体の標記パッケージ法について電氣的, 熱的, 機械的形質について検討し, 実際の応用分野での性能を述べ, 本方法の利点について述べた.

Subject packaging methods of power semiconductors are studied about electrical and thermal-mechanical properties, the performance in application fields the practice is described, and advantages of this method are described.

▶ プリント配線を使用目的で分類して各々の要求品質水準を示し, 品質管理上着目すべき箇所を列記.

Each required quality level is shown by classifying the printed wiring by use purposes, and the points are enumerated which should be noticed in respect of the quality control.

図-2 翻訳例

い場合には, 種々の可能性をさぐるので極端に長い時間を必要とする.

このシステムは, ソフトウェアの項でも述べたように, 大きな記述能力をもっているために, 今後生じてくる新しい言語現象に対してもかなりよく対応することができるものと考えている. したがって, 時間をかけて文法・辞書の双方を改良してゆくに従って, 翻訳できる言語表現の範囲は増加してゆき, システムの能力が高くなってゆくものと考えている.

当面の残されている課題は, 疑問文, 否定文などを含めて抄録文としては現れない文体に対して文法を適合させること, 代名詞, 省略などの文脈関係の文法規則をさらに充実させること, 電気工学以外の分野にま

で辞書を広げて翻訳文章の範囲を広げること、などであらう。

おわりに、本研究を国のプロジェクトとして採りあげ、計画の推進をしていただいた科学技術庁情報室、通産省工業技術院計画課の関係者諸氏に感謝する。この研究には非常に多くの方々への参加、協力をいただいた。特に草薙裕筑波大学教授、その他の大学関係者、元日本科学技術情報センター監事の中井浩氏、多くの民間企業の参加者の方々に感謝の意を表す。

参 考 文 献

- 1) 長尾 真：言語情報処理の過去・現在・将来，情報処理，Vol. 19, No. 2 (1978).
- 2) 長尾 真：科技庁機械翻訳プロジェクトの概要，情報処理学会，自然言語処理研究会資料38-2(July 1983).
- 3) 坂本義行：日本語形態素解析の基本設計，情報処理学会，自然言語処理研究会資料38-3 (July 1983).
- 4) 中村順一：文法記述用ソフトウェア GRADE，情報処理学会，自然言語処理研究会資料38-4(July 1983).
- 5) 辻井潤一：日本語構文解析，情報処理学会，自然言語処理研究会資料 38-5(July 1983).
- 6) 西田豊明：変換過程の基本設計，情報処理学会，自然言語処理研究会資料 38-6 (July 1983).
- 7) 中井 浩：語の収集と体言を中心とする辞書について，情報処理学会，自然言語処理研究会資料 38-7 (July 1983).
- 8) 坂本義行：格構造を中心とした用言と付属語辞書，情報処理学会，自然言語処理研究会，38-8 (July 1983).
- 9) 矢田光治，長尾 真：機械翻訳総合システムの基本設計，情報処理学会，自然言語処理研究会，38-9 (July 1983).
- 10) 中村順一：機械翻訳のソフトウェア環境，本特集号 (1985).
- 11) 辻井潤一：辞書の構成と機械翻訳，本特集号 (1985).
- 12) 辻井潤一，長尾 真：日英翻訳過程での処理とその翻訳結果への反映，情報処理学会，自然言語処理研究会資料，47-10 (Jan. 1985).
- 13) 長尾 真，辻井潤一，中村順一：日英科学技術文献の速報システムに関する研究の概要，情報管理，Vol. 28, No. 1 (Apr. 1985).
- 14) 長尾 真，辻井潤一：Mu プロジェクトにおける日英翻訳結果の評価，情報処理学会，自然言語処理研究会資料 47-11 (Jan. 1985).
- 15) 長尾 真：翻訳文の質の評価と言語の制限，本特集号 (1985).

(昭和60年5月16日受付)