機械翻訳とMuプロジェクト

坂　本　義　行（電子技術総合研究所）

現在、科学技術庁で行われています機械翻訳プロジェクト（Muプロジェクト）における機械翻訳システムについて御説明します。

機械翻訳は大きく分けると図1のような形で行われており、このような形が最も一般的な方式ではないかと思われます。翻訳にはいくつかのレベルがあり、まず形態素解析というのがあります。これは、たとえば日本語の場合は単語に区切る必要がありますので、単語に区切る操作を行います。単語から単語へ翻訳する単純な作業による翻訳は、われわれの研究所でも20年程前に行っていたことがあります。このような日本語の単語に区切ったものに対して、英語の単語に単に置き換えていくといったような語レベルの翻訳があります。それから、最近構文解析という言葉をお聞きになると思いますが、たとえば主語と動詞と目的語でできているような文を解析して、文のレベルで翻訳を行っておいて、今度はその文を組み合わせて生成していくという構文レベルの翻訳があります。さらに意味をも解析して翻訳を進める意味解析と、一つ一つの文だけではなくて人間の場合にストーリーというものがあるように、そのストーリーをも理解しないと本当の意味の翻訳はできないので、そういう文脈レベルの解析があります。

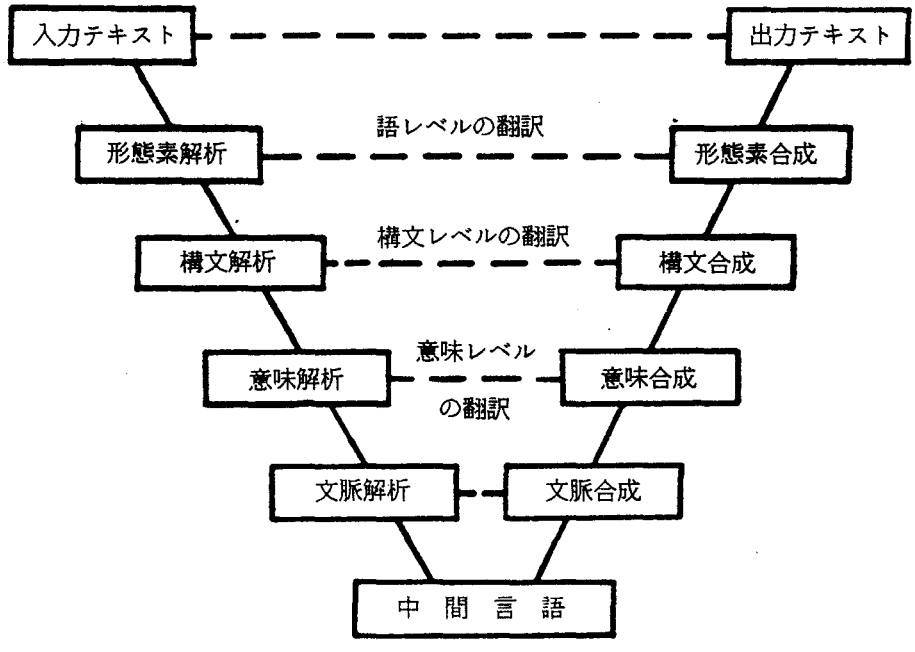

図. 1　自動翻訳システムのレベル

それから一般に、ピボットと呼ばれている中間言語があります。これはある言語で書かれた文章を一般的な構造まで解析して、あらゆる世界の言語に対して一つの言語体系で置き換えることができるといったレベルです。これが究極のもので、完成すれば全ての言語に適用できるので、いきなり日本語に翻訳されて出てくるというようなことができるわけですが、今のところは完成されていません。現在のところは構文解析のレベルから意味の解析のレベルへ若干入り込んだというようなレベルです。われわれの所で行っているのも、大体そのレベルの開発を進めていると考えていただいてよいと思います。

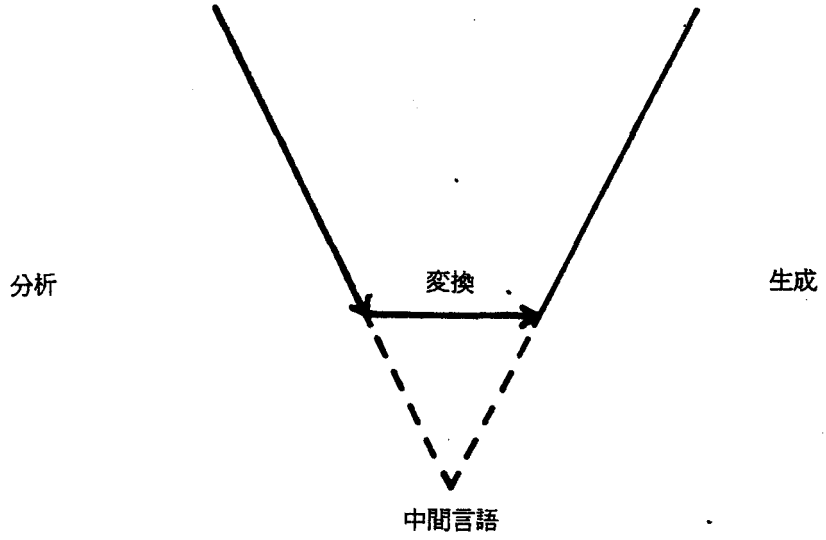

図. 2　トランスファー方式による翻訳

図2及び図3は、各レベルでどういうことを行っているかを表わしていますが、形態素解析は品詞や派生、活用あるいは慣用句を取り扱います。それから構文解析のところでは、その品詞の並びを見て、主語、述語の解析を行います。格支配は、一般に目的格とか所有格とかを、もっと複雑な形の格に分解して解析することです。それに対して語彙の変換は、日本語のある単語に対して、英語で何という単語に対応するかという置き換えを行うことで、そのためには当然、辞書、対訳語といった辞書が必要です。それから構文変換になりますと　「私は少年である」　というのに対して　「I am a boy」　というように動詞が前に来なければならないという英語の構造があります。そういう構造の置き換えを行わなければいけないわけです。この

ようなことを構文変換といいます。今度は英語の世界で逆に構文を英語のきれいな構文に直して、さらに、英語の中でたとえばedを付けたりingをつけたり、複数形の場合はsをつけたり、あるいはfootの時はfeetと置き換えたりする形態素合成を行います。そして目的となる英語が生成されるということになります。この方式をトランスファー方式といっています。

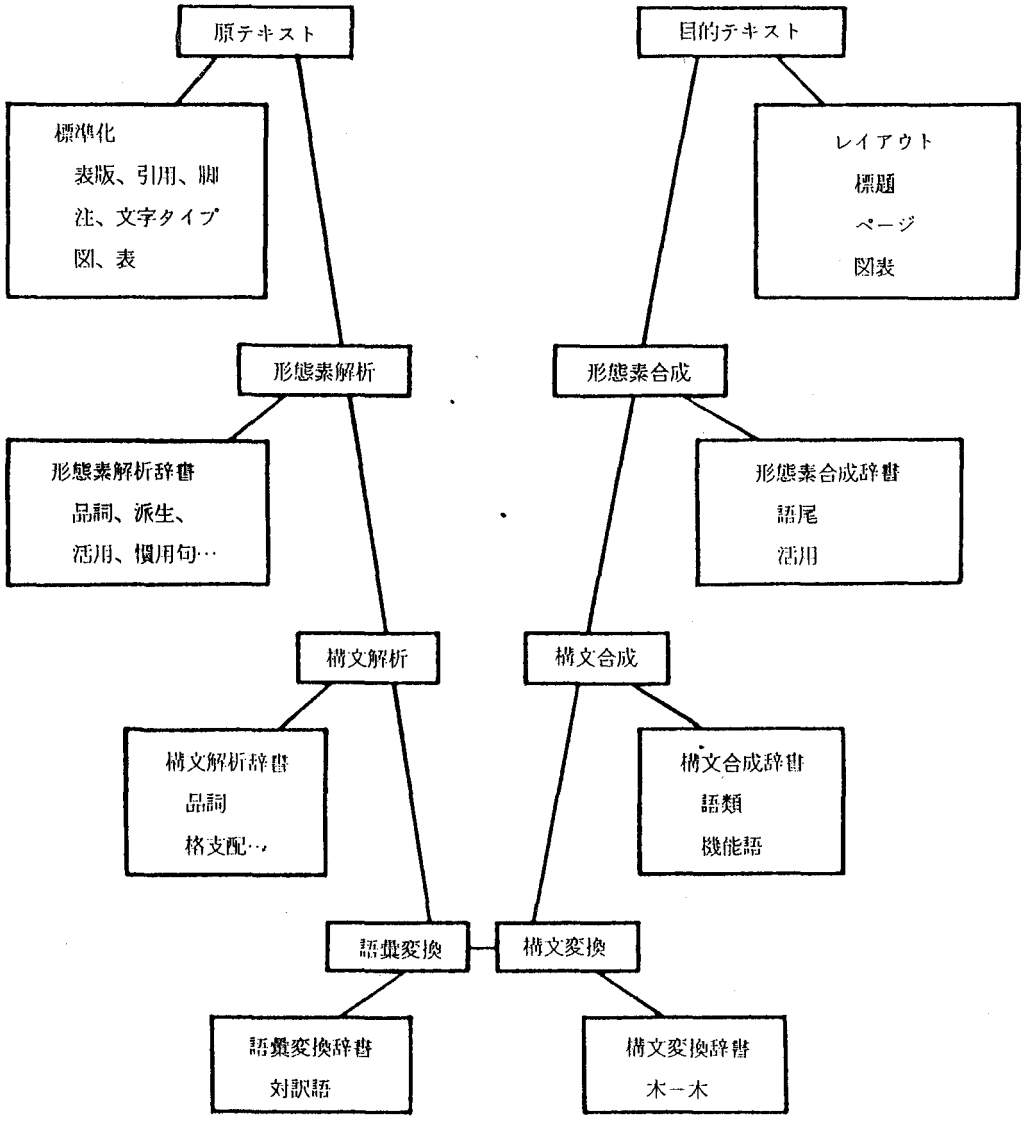

図. 3　トランスファー方式の構成図

現在、なぜ国のプロジェクトとして機械翻訳が行われているかといいますと、「科学技術は

ますます高度なものとなり、その知識、質、量共に加速度的に増大している。研究者がより創造的かつ効率的な研究活動を行うためには、広範な分野にわたる膨大な量の最新の科学技術情報を随時利用することが不可欠になっている」　これは皆さんが一番よく御存知だと思います。その中で2番目として　「わが国において、科学技術文献サービス、日本科学技術情報センター、通商産業省工業技術院情報計算センター、農林水産省農林水産研究情報センター、大学といった機関では、近年科学技術の高度化、研究活動の活発化により取り扱う文献の数が極めて膨大なものとなる。そして外国語で書かれた文献が、その中で７０%を占める様な比率になっている」　というわけです。学生の皆さんも外国文献を直接読まなければならないことが非常に多くなっていると思いますが、おそらく日本の研究者が読んでいる文献の半分以上は外国の文献だと思います。特に英語で書かれた文章は非常に多いということです。

「従来、科学技術文献は国内で利用するため、翻訳の需要が非常に強く、多くの文献が翻訳され利用されている。またこれらの文献は、データベース化する際には専門の翻訳技術者によりアブストラクション等の翻訳が行われている。しかしその量の増大により、こうした作業の機械化省力化が求められている」　ということが一つです。それから　「一方、わが国の産業活動、科学技術活動がますます高度化し国際化するに伴ない、わが国の科学技術情報は、ますます国際的に利用されつつあり、科学技術情報活動の国際化が必須の問題となってきている」ということがあります。さらに　「近年わが国の高度な科学技術水準に注目した欧米先進国や、わが国が研究協力・技術協力などで重要な役割を演じている発展途上国などから、わが国の科学技術文献の提供が強く要請されているが、言語が障害となって充分これに対応できない状況にあり、対応策が求められている」　すなわち、日本語は非常に特異な言語で、特に欧米では日本語を理解する人は非常に少なく、日本語の文献は最近彼等を手こずらせています。彼等がそれを理解することは非常に困難であるがゆえに英語になった文献が欲しいという要請があります。

「こうした問題を解決するため、近年進歩の著しい情報技術を活用した日英科学技術文献の速報システムの実現が望まれる」という理由から、科学技術文献アブストラクトの翻訳を支援することがねらいです。われわれのプロジェクトの研究項目は、最終的には日英科学技術文献の速報システムの実現ですが、英語は非常にあいまいな表層の構造を持っているために、機械で翻訳することは容易ではありません。日英翻訳をとりあげたのは、需要の面からです。

次に科学技術文献の翻訳について御説明します。当然、川端康成の小説なども翻訳してみたいという希望はありますが、小説ということになれば御存知のように、その意味の多義性というか、深いニュアンスまでも訳出しなければなりません。これは今の段階では非常に困難な部分を含んでいるということから科学技術文献に限定しました。それから速報システムですが、文献の内容を簡単に理解するには、まずアブストラクトを見ます。また、これを出来るだけ早

く読みたいという要望がありますので、その速報を翻訳することが非常に大切と考えて、Muプロジェクトで日英の科学技術文献の速報に対する翻訳システムを作ろうということから、われわれの研究がスタートしました。

この翻訳システムは、京都大学の長尾先生の機械翻訳方式をベースにして、汎用的な翻訳システムというものを構築しようと考えています。このシステムの特徴は、まず第一に、翻訳のメカニズムの基本操作として、TreeのリストからTreeのリストへのパターン変換の機能をとります。これは将来現われてくると思われる相当複雑な言語理論にも対処できる能力を持たせております。Treeは木構造と呼ばれており、要するに1つのセンテンスを木の構造に解析するわけです。その木の構造を別の言語の木の構造に置き換えます。例えば日本語と英語の木の構造ですが、先程の 「I am a boy」 でも、その構造はいつも変わっています。そういう構造を木の形で与える。そうすると日本語の木から英語の木へ置き換えます。そういうパターンからパターンへの置き換えだけで処理するようなシステムを作っておきます。そうすれば、たとえば日本語からフランス語、あるいは英語からフランス語といろいろな言語に対しても、パターンからパターンに変換して翻訳を行えば非常にスムーズに行くし、汎用的なシステムとして作りあげられるだろうということから、Treeの変換方式をとろうということになりました。次に言語情報を記述するためのわかりやすい記述システムをつくります。これは計算機のことを知らせたい人達、特に言語学者に文法辞書等の作成をしてもらえることになります。しかし機械翻訳というのは、もちろんコンピュータを知らないと開発はできないわけですが、実際には辞書を作ったり文法を作ったりする時に、言語学者あるいは心理学者、哲学者といったような計算機を知らない人達の援助も欲しいので、そういう人達も使えるような翻訳システムにしておきたいわけです。それからプログラミング言語として文字列を置き換えるのに非常に便利なLISPを採用しました。4番目は先程申し上げたようにトランスファー方式であるということです。5番目に、解析は格文法を中心として意味の取り扱いを重視します。日本語を取り扱う時には、意味を中心とした格文法の考え方が、現時点で最もよいという理由からです。格文法というのは簡単にいうとテニヲハです。何々が、何々を、何々に、どうどうした、というような言い方です。英語は順序が固定しているので構造の文法で解析すると便利ですが、格の組み合わせでセンテンスが成り立つ日本語の文法解析には格文法が一番適していると言われています。6番目に辞書情報を中心に処理を行います。これは多くの特殊な言語現象も取り扱えるようにするために重要な概念です。いわゆる言葉を扱うわけで、その知識ベースになる部分というのは言葉に対する辞書ですから、最近特にその重要性が認識されてきています。辞書を置き換えればさまざまな言語の翻訳もできるし、さまざまな形の文章をも翻訳することができます。それは全く独立し、切り離された辞書をつなぐことによって使用できるので、切り離して作っているわけです。

以上がMuプロジェクトで使うトランスファー方式の特長です。簡単な処理手順を図4に示しましたが、これは図3を逆向きに表わしただけで、形態素解析、構文解析、それから構文変換、構文合成、形態素合成という一回りになります。中心にあるのがこれを翻訳するための核となるソフトウェアです。逆にこの外側にいろいろな文法や辞書をぶら下げています。電気、計算機、化学などの各分野の翻訳をする場合は、当然使用される単語も異なるので、それに対応するための専門用語データベースを構築することになります。

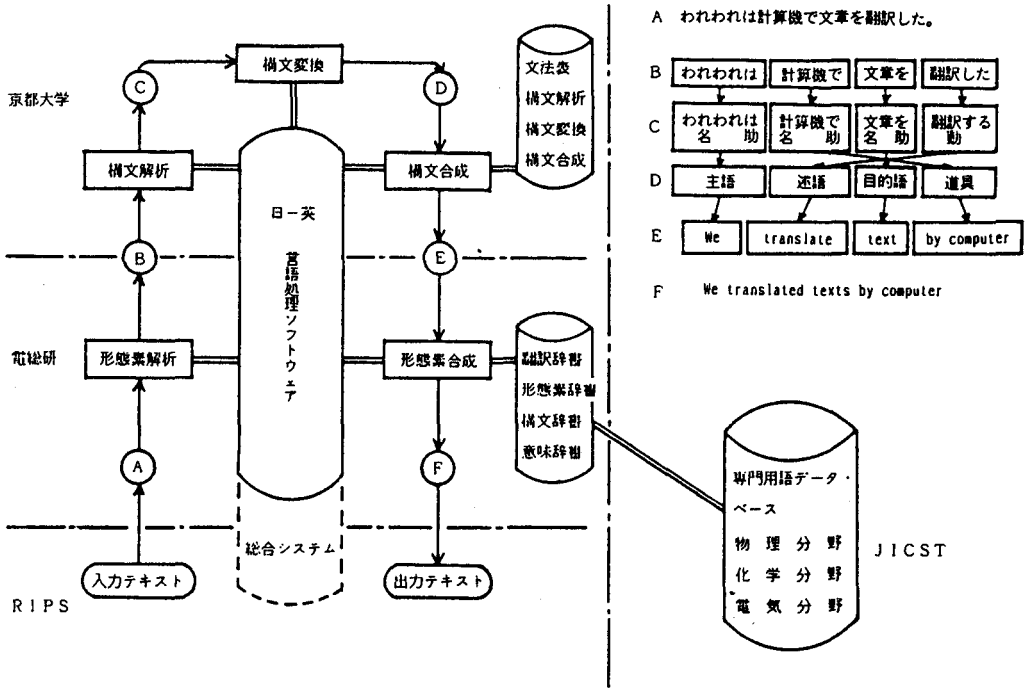

図. 4 文献翻訳システム概念図

図4の右上にあるのが一つの例ですが、「われわれは計算機で文章を翻訳した」という文章の場合は、形態素解析で図のように単語に分割できます。この単語に対して品詞が与えられていますが「われわれ」は名詞で、「は」は助詞であり、「計算機で」は名詞と助詞です。「文章を」は名詞と助詞であり、「翻訳した」は動詞である。ここで、この「は」、「で」、「を」が格構造に変わり、主格、述格、目的格、道具格に置き換えられて、順序も置き換えられています。そして「We translate text by comuputer.」という英語に変換しているわけです。も

ちろん、もっと複雑な文章の場合には、構文合成というところで複雑な処理が行われます。最後に形態素処理が行われると「翻訳した」という過去形になっていますから、translatedとedがつけられています。また、「text」は一つのtextでなくて「文章を」だから、sつける処理を行います。これが非常に簡単にいって翻訳のメカニズムというわけです。そしてその過程にさまざまな問題がたくさんあるわけです。

たとえば辞書をつくる時、同じ言葉でもいくつかの意味をもっているので、細かい記述を必要とします。そのために文法としては格文法をとることにし、表－１に示す３３個の格を採用していますので、１つの動詞の辞書をつくるのにも非常に時間を要します。

	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－１．　日本語格ラベル一覧表

また言語学者といいますか、言語の専門家も必要になります。それから言葉の表現は一様ではなくて、いろいろ言い方ができますから、辞書を作っていく時に何らかのマニュアルが必要になります。さらに普通の文章の中には能動的な文章のほかに、受身で書かれた文章があります。そのような文章を等価な能動の態に置き換えることを行わないと、実際に辞書をつくる時の情報が得られません。１つの動詞にもいくつかの意味がありますから、いろいろな名詞に対して分類したコードを与えておいて、動詞との組み合わせを辞書の中にきちっと記述しなければなりませんので、動詞の辞書は非常に多くのことが記述されることになります。われわれのところでは現在、約３０００語の動詞の辞書を作っていますが、動詞の辞書をいかにきちっと作るかということが、翻訳の質を本質的に決めることになります。次に、実際に翻訳の過程でどういうことを行っているかということですが、最初単語に切って形態素解析を行い、次に構文解析に入ります。１つの文に対して、たくさんの解析が行われるわけですが、その中でどれが妥当であるかというのは非常に難しい判断があるわけです。今のところ、それが本当の意味で妥当であるかどうかは別にして、一番妥当だと思われる解を１つだけ出しています。それから、単語個別の規則をその単語の辞書規則として登録しておくことができるため、全体の文法系を大きく変更することなく、個別的な現象に対処することができます。だから単語単語に辞書規則を書けるように、言語的な情報のほかに文法的な記録も書き込めるようになっています。そのことによって、部分的な置き換えだけで使うことができる特徴をもっています。今実際に翻訳の対象としているのは抄録文であり、実際にはＪＩＣＳＴ発行の科学技術文献抄録について開発を行っています。文法は現在２０００以上もありますけれども、それをただ並列に並べているのではなく、まとめたものを使っています。それでその置き換えによって他のテキストに対しては、サブグラマーを用いるといったことも図られています。変換過程についても、日本語といった特有なものから英語へと置き換えるのは非常に変わっているわけですから，難しい問題がたくさんあります。

変換が行われたら英語の世界に入りますが、今度は英語の世界での独特の部分があります。例えばIt...that構文や、名詞が２つ並ぶとき片一方を省略するとか、ｓのつけ方等や、冠詞とか英語特有のものがあるわけです。

現在のところは、やっと日本語から英語に変換するといいますか、翻訳する過程が開発終了しつつある状態です。同時に英日の翻訳の開発も進めています。日本語から英語へ変換するためには、日本語の辞書、日本語から英語へ移す辞書、英語の辞書を作っているわけですけれども、英日の時には英語の辞書、英日の辞書、日本語の辞書が必要になります。現在のところ日本語や英語の辞書はどちらの方向にも使えるように設計をすすめていますが、変換用の辞書は逆方向にそのまま使用するわけにはいかないので、日英と英日のものは、別のものを使用する形でシステムが作られています。今までの話は、いわゆる翻訳の核の部分です。このＭｕシス

テムは６１年の３月まで開発が進められ、その段階で一応、計算センターに導入して、研究者に自分の論文などを翻訳させるのに使ってもらおうという試みがありますので、使いやすいものにしておかなければなりません。

総合システムの簡単な思想だけお話ししておきますと、２種類の翻訳のやり方に使いたいということです。１つは多量文章の一括翻訳、計算機を使われる方は御存知だと思いますが、バッチ処理と同じで翻訳したい文章の磁気テープを作り、それを渡すと翻訳をきちんとやり、その磁気テープを返してくれるといったような形のものです。もう１つは論文作成翻訳で、研究者が自分の端末で日本語を入れるとそのセンテンスの翻訳文がすぐに出てきて、それを自分が気に入らなければその場で直し、編集することができるものです。原文と訳文の同一編集というのは１つの画面の上に原文を入れると、その訳文が画面の下半分に出てきて、その段階で出力の訳文が自分の求めるものでなかったりすると、部分的に修正することができます。あるいは、日本語が原因で英語がまずいときには、日本語の方をもう少しきれいにするといった編集もできるようにしておきます。そのような編集をするためには日本語用、英語用のエディターが必要になります。それから、現在計算センター等ではどこでもありますが、いろいろな文献をサービス形態として見ることができます。抄録などを検索しますと自分の欲しいものに関する文献が現われますが，それが英語の場合、日本語で見たいと思うことがあります。そこで、文献検索システムと結合して翻訳をすると、英語の文献が日本語になって出てくるような使い方もできるようにしておきたいところです。また、辞書の方がまちがっていたときの辞書の編集とその管理、化学あるいは土木といったような種類の文章に合わせた辞書の選択もできるようにしておきたいところです。得られた翻訳テキストはどういう形で使われるのかわかりません。プリントアウトしたものが欲しいのか、翻訳結果が良くなくて、翻訳者にエディッティングしてもらいたいから磁気テープの形で出したいといったこともあります。だんだん翻訳が多くできてくると管理を必要としますから、管理機能を持っていなければなりません。実際にはこういったことを考えてサービスをしようという話になっています。

以上のように、京都大学などと協力してＭｕプロジェクトという１つの翻訳システムの開発を行っています。今は日英だけですから英日も行わないといけませんし、問題はたくさんあります。たとえばテキストは、電気工学の分野の中のＪＩＣＳＴで作っている抄録ですからテキストが変わったらどうか、また他の言語への拡張能力を持っているのか、翻訳したものが本当に正しい英文なのかどうかというようなことです。

今まで機械翻訳そのものは何十年間か行われてきているわけですけれども、一番困っているのは翻訳された結果の質の評価が非常に難しく、誰も定量的に評価できないので、基準をきちっと定められないということです。そこで今度は逆に、翻訳する過程において、どこでどういう間違いをおかしたかということがきちっと整理されれば、それが翻訳の評価にもつながるだ

ろうと考えています。人間が翻訳する時には専門の人が翻訳するわけで、その人のキャリアで行ってしまうわけですが、キャリアとは何だということです。日本語をどれくらい理解できて、英語にどれくらいきちっと置き直して、さらにその置き直された英語をきちっといわゆるnative speakerが理解できるような英語に置き直す・・・とそういう過程があるわけです。そういう過程を、逆に、きちっとしてやれば、その質の評価も自づからできるようになってくるだろうという問題があるわけです。ですから、そういう意味では１つの試みを行ったわけで、最初に申し上げたように、機械翻訳がものすごく進んだという話では決してありません。

要するにごく一部がとりかかり、そして計算機の上で、一般の人が翻訳というものを使うことができるレベルに達してきたということであって、翻訳者がいらなくなるという話では全然ない、ということを理解していただければいいのではないかと思います。

（昭和５９年１２月６日講演）