

02-DPC-09

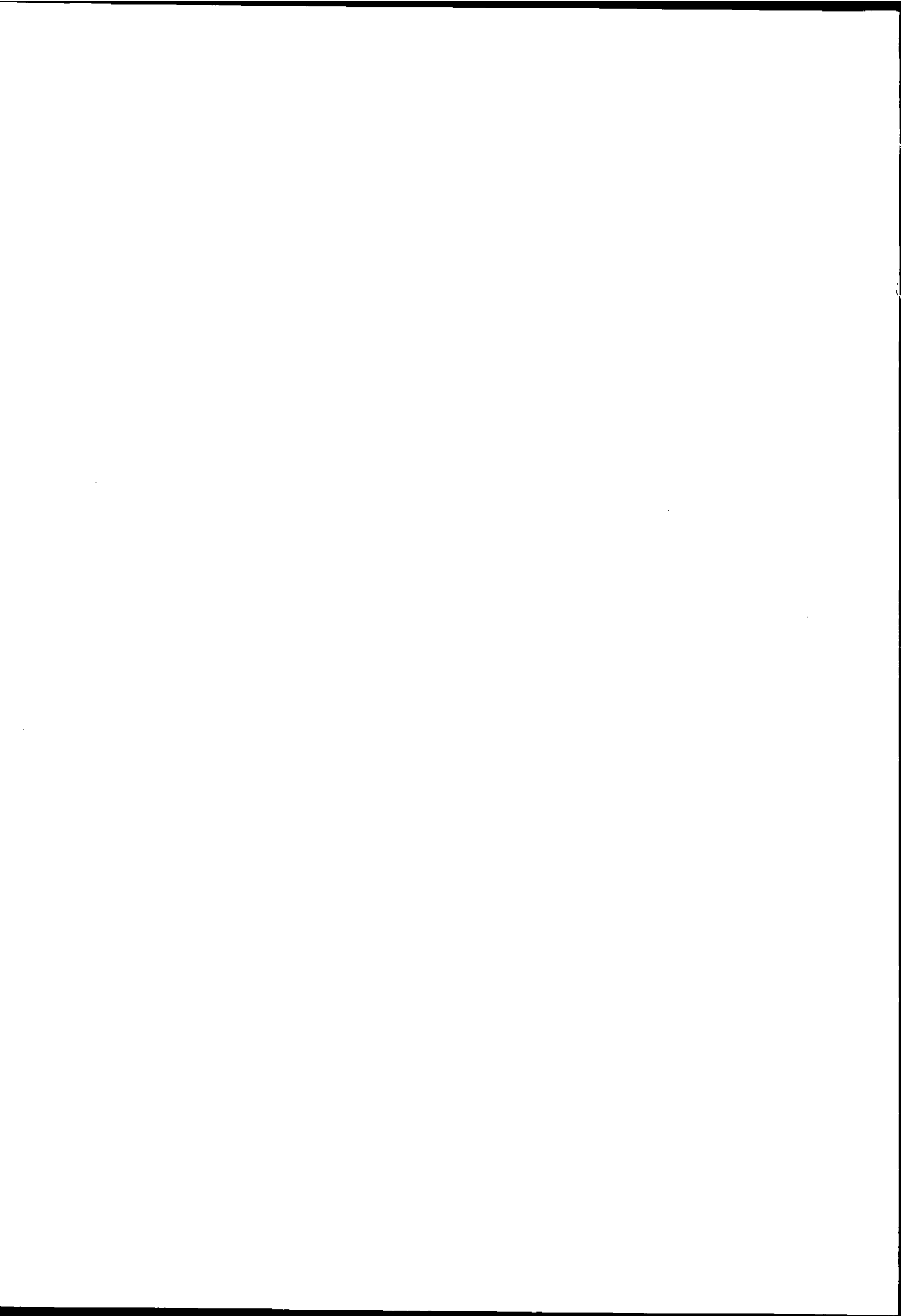
保存本

国際問題委員会報告書

平成3年3月

財団法人 データベース振興センター

本報告書は、日本自転車振興会から競輪収益の一部である機械工業振興資金の補助を受けて平成2年度補助事業の一環として作成したものである。



はじめに

わが国の国際的地位の向上とともに、日本の科学技術情報およびビジネス情報に対する海外からの関心は高まる一方です。今日、欧米の企業にとって経営戦略を展開するにあたり、日本情報は必要不可欠なものとなってきています。

日本情報の利用にあたって、外国からみて最大の問題は言語であります。それと同時に、日本国内のどこでどのような情報が発生しているのか、またその情報をどうすれば入手できるのかが分からないという不満も多くあります。このような海外からの要望に応え、日本情報の国際的流通を促進することは、経済大国日本の義務でありましょう。

(財) データベース振興センターは平成元年度に国際問題委員会を設置して、日本情報の国際化を促進するにあたって解決されなければならない問題点を整理、検討し、提言を行って参りました。その主な内容はデータベースの英文化（機械翻訳の促進）、海外ニーズの正しい把握、定例的広報宣伝活動、常設的データベース案内機関の設置、海外提供の支援施策・助成制度等です。この提言を受けて平成2年度の国際問題委員会は、日本情報の海外提供にとって大きな障害となっている言語、すなわち英文化の問題に焦点を絞って、データベースに関する機械翻訳システムの対応について検討を進めて参りました。

本報告書は「総論（わが国の機械翻訳の現状と問題点）」、「機械翻訳システムの開発と現状」、「機械翻訳システムの周辺の現状」、「機械翻訳の利用の現状」の4部から構成され、機械翻訳の技術的側面と実務的側面からとりまとめました。広く関係各位の参考になれば幸いです。

最後に、本報告書を取りまとめるにあたりご協力をいただきました委員の方々をはじめ、通商産業省など関係各機関の皆様方に心より感謝の意を表します。

平成3年3月

財団法人データベース振興センター
理事長 圓城寺 次郎

「国際問題委員会」委員名簿

委員長 根 岸 正 光 学術情報センター 教授

委 員 大 熊 喜 宣 (株)平和情報センター データベースサービス部

柿 元 俊 博 富士通(株) 企画部A I 開発推進部機械翻訳課長

神 尾 達 夫 日本経済新聞社 データバンク局記事情報部長

空 閑 茂 起 シャープ(株) 技術本部情報技術開発センター第2 開発室長

高 浜 忠 彦 リソースシェアリング(株) 代表取締役社長

富 永 勲 (特)日本科学技術情報センター 情報部英文データベース部門 主任情報員

宮 澤 利 成 日本貿易振興会 企画部情報システム室長

モリス・ブレン 駐日E C委員会代表部

事務局 井 出 真 弘 財団法人データベース振興センター企画部長

佐 野 英 之 財団法人データベース振興センター企画部次長

小 泉 幸 一 財団法人データベース振興センター企画

目 次

第1部 総 論

わが国の機械翻訳の現状と問題点	1
-----------------------	---

第2部 機械翻訳システムの開発と現状

第1章 日本科学技術情報センター(JICST)における機械翻訳システムの開発と現状	7
第2章 富士通の機械翻訳システムの開発と現状	15
第3章 シャープの機械翻訳システムの開発と現状	21
第4章 RSC & CLI における機械翻訳システムの現状	39

第3部 機械翻訳システムの周辺の現状

第1章 学術用語集のCD-ROM化	49
第2章 シソーラスの開発と利用	55

第4部 機械翻訳の利用の現状

第1章 G-SEARCH の機械翻訳による海外提供	65
第2章 日本貿易振興会(ジェトロ)における機械翻訳利用の現状	71
第3章 EC 委員会における機械翻訳システムの利用	75

第1部 総

論

わが国の機械翻訳の現状と問題点

わが国の機械翻訳の現状と問題点

1. はじめに

(財) データベース振興センターの国際問題委員会では、平成元年度において、「日本情報の国際化」、すなわち国産データベースの海外提供の推進をテーマとして検討を行い、これに向けての提言をとりまとめた。そこでの結論の重要な1項として、国産和文のデータベースの外国化（とりあえずは英文化）の問題がとり上げられ、これに対して機械翻訳の適用可能性を今後一層追求するべしとの結論に至っている。この点を受けて、平成2年度の国際問題委員会では、データベースにおける機械翻訳の可能性を中心として検討することとし、これに関する専門家に委員を委嘱して、技術的側面とデータベースの海外サービスの実務的側面の両面から検討を行った。以下、本委員会における検討の結果を総括的に示すことにする。

2. データベース・サービスにおける英文化の方式

和文のデータベースを英文化して、海外利用者に使い易くしようとする場合、単純には、データベースの全内容を予め英文化した後に、本来の英文データベースと同等に扱ってゆくというのが、最も分かりやすい方式である。しかし、こうした本格的英文化は、実際には最も経費と時間を要する方式であって、和文データベースの海外での利用促進の観点からは、もう少し簡便、低コストでしかも一定の効果をもたらす方式もありうる。これらを含めて整理してゆくとつぎのようになる。

2-1 翻訳の対象項目

(1) メッセージ、項目見出し、定型的項目のみの英文化

英文化の対象とする項目としては、まず検索コマンドの入力を促し、また検索結果のヒット・レコード数などを通知するメッセージの類がある。しかし、これは、メッセージ表を和英文で切り換えるという方法で、多くの検索システムですでに実現されており、特に問題はない。次の段階として、検索結果表示の際の項目見出しと、日付とか分類など内容が定型的な項目に関するもの、これらを英文化して表示するものが考えられる。これは翻訳というより、単純な変換というべきものであるが、特許データベースについて、端末側検索ソフトウェアに、こうした変換機能を組み込んだものが海外で実用されて、一定の効果をおさめている。

(2) キーワードの英文化——和英対訳辞書・シソーラス

検索の際に利用者が指定するキーワードに英文が使用できることは、データベース英文化の第1段階として重要である。検索の手掛りとするキーワードは、原則的に利用者側で独自にこれを

想起して指定する必要があるもので、これが英文で可能になることは外国人利用者には大きなメリットになるものと思われる。

科学技術論文などでは、著者の段階で標題、著者名、所属機関、キーワードについて英文標記を並列することはすでに随分普及しており、英文要旨の添付も進んでいる。つまり、こうした分野ではキーワードの英文化はすでに相当程度行われているとみてよいであろう。

機械翻訳の立場からみると、キーワードの翻訳は、文法解析を要しない、むしろ変換というレベルのものであるから、その限りで技術的な問題は少ない。つまり原理的には、和英対訳の辞書、用語集の類があれば足りるということになるのであるが、実際は、この種の辞書の整備と的確な更新が大きな問題になる。進歩の著しい科学技術分野や時事的分野の場合、キーワード辞書の更新は容易ではない。

安達講師の講演では、学術用語での対訳辞書整備の実務的問題点が指摘されたが、この場合、標準的な用語の設定という規格制定的な側面が加味されており、辞書整備やその更新はなおさらずかしいことになる。

もっぱらデータベース検索を目的とする辞書の整備に関しては、神尾委員から「日経シソーラス」に関する報告があった。新聞記事データベースの場合、同一対象について種々の表現が用いられることが多いから、シソーラスによる用語の統制と包含関係の規定は検索もれを防ぐのに有効である。しかしこの事例でも、新語の出現を常時監視しながら、3年程度の間隔で定期更新を行うために、専門の担当者を複数名配置するなど、シソーラスの維持更新は相当の業務量になることが指摘されている。なお、この例は現状では日本語のシソーラスであるが、こうしてシソーラスが維持されるならば、これに英訳を付加することは、業務量的にさほど問題にはならないと見込まれるとのことである。つまり、「原文」である日本語辞書の維持管理の方が大いに問題であるということであり、これが確保される場合には、その英文化は十分現実可能であろう。

なお、ディレクトリー、統計類のデータベースで、基本的に文章を含まないものの英文化は、全面的に上記対訳辞書に基づく変換方式によることになる。

(3) 本文全体の英文化

データベースの本格的な英文化は、本文を含む全項目を英文化することになるが、このとき本文とは抄録型データベースでは抄録であり、全文データベースの場合はさらに本文全文の翻訳を含む。以下においては、主として抄録型データベースを念頭にして、その翻訳における機械翻訳の適用可能性を検討することにする。

2-2 翻訳の実施者と方法

(1) 著者翻訳

翻訳の方法には、従来からの人手によるものと機械翻訳システムによるものがある。まず著者自身による翻訳については、少なくとも科学技術系の邦文では、前述のとおり、標題、キーワー

ドに英訳をつけることはすでに普及しており、さらに英文抄録もかなり普及しているとみられる。もっとも、これはいわゆるフル・ペーパーに関してであって、学会報告その他、グレー・リテラチャーと称される種類の文献では、こうした英文化はそれほど進んでいないのが実情であろう。学術情報センターの学会発表データベースでは、著者英文抄録の付加割合は30%程度である。近年、海外からの関心は、わが国で進行中の現況を知るものとして、こうした流通性の低いものにも集まっている。

ブレン委員の報告にある、ECによる日本文献情報の収集事業は、まさにこの線に即したもので、また機械翻訳の先駆的適用事例としても重要である。

宮澤委員によれば、JETOROのような元来海外向きの組織では、外国語出版物の制作が従来から定常的に行われているが、その作成方法は著者翻訳の一種と考えられる。そして、今後翻訳量を一層増大させるためには、作業能率の向上のため、機械翻訳に期待するところが大きいとされる。

機械翻訳システムは、空閑委員、高浜委員の報告にあるように、パソコンやワーク・ステーションなど、個人利用型の機器で単独に稼働するものの開発が進んでいる。これらは簡便に導入できることを目指しており、著者側での英文原稿作成に役立ち、英文文書の量的拡大に資するものと思われる。

(2) データベース編集者側人手翻訳

データベース編集者側での翻訳専門家による翻訳は、これまでの英文化データベース作成において主流をなしてきた方法である。しかし、これには人員的な限界があつて、今後英文化を飛躍的に拡大するには機械翻訳システムの活用が不可欠と考えられる。

(3) データベース編集者側機械翻訳

データベース編集者における機械翻訳の適用には、大量のデータを定常的に翻訳してゆくという運用体制を含めた「システム化」にも大きな課題がある。

富永委員による JICST での機械翻訳システムの適用経過報告は、データベース編集者側での、全面人手翻訳から、機械翻訳を併用した方式への移行を具体的に解説しており、今後の方向を示唆するものであろう。

(4) データベース利用者側翻訳

利用者における翻訳についても、今後はこれに機械翻訳を活用してゆくことが可能であろう。パソコンやワーク・ステーションなどで稼働する機械翻訳システムの海外での普及が図られれば、海外のデータベース利用者側での翻訳に大いに貢献すると思われる。とくに、全文データベースの場合、本文全文をデータベース提供者側で翻訳して提供することもありうるとはいえ、利用者側で機械を利用して翻訳するという方法も、大いに有効であると思われる。

2-3 翻訳の時点と対象レコード範囲

(1) 事前に全レコードを翻訳

翻訳をデータベース・サービスのどの時点で行うか、またレコード全件について翻訳するか否かについて検討すると、単純には、英文データベースとしてサービスするからには、予めレコード全件に英訳を与えたものを、サービスに供するというのが自然である。

全件事前翻訳では翻訳に要する費用が問題になる。一般にデータベースに含まれる大量のレコードの内、実際に検索出力されるのはごく一部であるといわれる。事前全件翻訳の場合、こうした結局使われないレコードについても翻訳を施しておくことになるから、英文化コストは大きくなる。機械翻訳の適用による、翻訳コストの低減が大いに期待されるところであろう。

なお、この方法では、専門家による校正を併用して翻訳の品質向上をはかることができるし、また英文データベースとして、磁気テープなどで他のベンダーに頒布することもできるといった利点がある。

(2) 検索結果表示の際に当該レコードのみ翻訳(オンデマンド・リアルタイム翻訳)

高速な機械翻訳の現実に伴って、あらたな可能性が生まれつつある。すなわち、検索の結果としてレコードが表示される際に、機械翻訳システムを介して、これを英訳して出力する方法である。この場合、人手をまったく介さないから、翻訳コストは低く抑えられる。

大熊委員、柿元委員によれば、この種のオンデマンド・リアルタイム型機械翻訳によるデータベース・サービスを実験して、良好な結果が得られており、今後その実用化を追求しつつあるとのことである。

もつとも、このように機械翻訳に全面的に依存するものとすれば、全件事前翻訳を行ってもさほどのコスト要因にはならないという面もあるから、いずれにしても機械翻訳実用化の重大性が指摘できる。

2-4 機械翻訳の実行サイト

(1) ホスト機側オンライン/オフライン——プロデューサー、ベンダー、ゲートウェイ

機械翻訳をどこで行うかを考えると、一般にはデータベースの提供者側でこれを行うことになる。事前翻訳の場合はもちろんこれに該当するが、オンデマンド型機械翻訳のときも、翻訳システムはホスト側コンピュータに搭載されることになる。ただしこの場合は、ベンダーあるいはゲートウェイのコンピュータ上で稼働することになるので、英文化はベンダーやゲートウェイ独自の付加価値サービスという位置付けになる。この点は、今後わが国におけるデータベース業界の構成、役割分担との関連においても注目する必要がある。

(2) 端末機側オンライン/オフライン

機械翻訳システムの進歩とパソコンあるいはワークステーションの高性能化に伴って、あらた

な可能性が考えられる。すなわち、翻訳システムを端末側に搭載して、和文で送られてきたレコードを利用者側で翻訳して表示する方式である。

そもそも利用者側に翻訳システムがあれば、検索結果をファイルに保存して（ダウン・ロード）、これを翻訳システムにかければよく、これはデータベースの検索とは別個の作業として可能である。しかし、本格的には、検索機能と連動した翻訳システム、すなわち翻訳機能内蔵型の検索ソフトウェアが必要である。

現在、データベース検索のためには、特定のデータベースやベンダー・システムに依存しない、汎用の通信ソフトウェアが使われているが、上記のような翻訳機能内蔵型検索ソフトウェアで汎用のものの実現は難しいとみられ、データベースやベンダー・システムに個別に対応したものが必要になるであろう。

英文化データベースの普及という観点では、ごく一般的な英文の通信ソフトウェアが使えることが有利である。してみると、オンデマンド型翻訳の場合も翻訳システムはホスト側に搭載して、事前翻訳型と同様、利用者からは普通の英文データベースであるかのようにみえるのがよいと考えられる。

3. まとめ——データベース・サービスとしての評価基準

上述のような諸側面からの検討を踏まえて、和文データベースの英文化サービスを考えるとき、当面現実的なものとみられる次の2つの方法を中心に、今後さらに検討してゆくのがよいと思われる。第1は英文による検索+和文出力というアクセス側のみの英文化であり、第2は英文化による検索と出力というアクセスと出力両側の英文化である。

3-1 対訳辞書による英文化アクセス——受動的日本語能力

和英対訳辞書・シソーラスを介して、英文による検索を可能とするが、抄録などの出力は和文のままとする方式は、一見中途半端な感じを与えるが、これは海外利用者側における日本語能力の問題に関連して評価されるべきである。外国語の読み書きでは、一般に読めるための受動的語学力よりも書けるための能動的語学力の方が相当難しい。これをデータベース検索に当てはめれば、検索のためのキーワードを想起して指定する部分が書く能力に当たる。従って、出力は和文でも、この検索部分が英語でよいということになれば、利用者側に要求される日本語力は随分ゆるやかになり、外国利用者層の拡大に対する効果は大きいと考えられる。さらに、かな漢字変換を伴う日本語入力も、オンライン検索では持続時間と関連して大きな問題で、これをすばやく行うのは容易ではなく、この点でも、英語による検索の有利性が指摘できる。

このように、アクセス側の英文化は、英文化データベース提供の第1段階として十分有効であり、このためのシステムと和英対訳辞書・シソーラスの開発が推進されるべきであろう。

3-2 機械翻訳による英文化データベース——機械翻訳の「実用性」と評価基準

データベースへの機械翻訳の適用に関しては、翻訳結果の品質に対する要求、考え方が基本的な問題であろう。これまで、機械翻訳の品質水準は、専門翻訳者による翻訳と対比する形で、その実用性が問われてきた。そして現状では、下訳程度にはともかく使えるという程度に機械翻訳のレベルは向上してきている。ここでさらに一層の品質向上をめざして、研究開発が進められるべきはもちろんである。しかし、その一方において、現状レベル程度のシステムを前提とした上で、データベース英文化のために、これをどのように使いこなしてゆくかという点で真剣に追究すべき段階に至っていることも事実である。すなわち、もはや使いものになる、ならないという議論の段階ではなくなっているといってもよいかも知れない。

もともと、文献抄録型データベースの場合、キーワードで検索して、抄録で内容を判別するという便宜的な方法、あるいは文献本体を逐一点検するのに比べるならば粗略ともいべき検索方法が採られている。キーワードといい抄録といい、これらは要するに検索手段としての文献内容の写影であるとみれば、その品質は検索効率との関連で評価されるべきである。そこで英文化データベースについて、これを平たくいうなら、英訳として正しいかどうかといった問題ではなく、利用者が、分かろうと思って読めば、何とか文献内容の判別できる程度の翻訳になっているかどうかという点が評価のポイントになるはずである。この点で、データベースは、一般の翻訳物とは大いに目的を異にし、従ってその品質評価基準も自ずと異なってくる。

こうした見方からすれば、現状レベルの機械翻訳は、データベースの英文化に対してすでに相当程度の実用化を獲得していると評価することができるであろう。従って、これをデータベースの英文化にいかに関活用してゆくかという観点からの議論を、今後一層積極的に展開してゆく必要があると思われる。

3-3 海外需要の動向

ところで、上記の英文化アクセス方式や機械翻訳データベースに関して、海外利用者側の反応が重大な問題である。すなわち海外利用者において、上述のような、何とか分かればよい式の評価基準、立場がどの程度存在するかということである。日本情報に関して、この程度の障害、欠陥を十分乗り越える程の強い需要がどの程度存在するかということである。むろん完全な英文データベースでなければ使いものにならないといった議論も依然有力であるかも知れないが、需要と供給の間での相互作用、すなわち両者におけるにわとりとたまごの関係を思えば、供給側として可能なことは、ともあれこれを展開してゆくことが肝要であろう。この視点から、英文化アクセス方式のデータベース・サービスや機械翻訳データベースの実用、拡大は推進されるべきものであろうし、これに対して応分の公的な支援策も講じられるべきものと思われる。

第2部 機械翻訳システムの開発と現状

第1章 日本科学技術情報センター (JICST) における 機械翻訳システムの開発と現状

第2章 富士通の機械翻訳システムの開発と現状

第3章 シャープの機械翻訳システムの開発と現状

第4章 RSC & CLI における機械翻訳システムの現状

第1章 日本科学技術情報センター（JICST）における 機械翻訳システムの開発と現状

1. JICST における開発の概要

1-1 開発の背景

JICST における科学技術情報の海外提供の現状は、平成元年度活動として当委員会がまとめた報告書「日本情報の国際化」（1990. 3）に述べてあるためここでは省略するが、最も新しい情報を紹介すると、平成2年度では約235,500件の技術論文を英文データベース化し、従って、1985年からの累積では約140万件の日本情報を海外向けに提供していることになる。

平成元年度報告にもあるとおり、英文データベースを作成する際の最大の問題点は「日英翻訳」であり、海外から要求の強い抄録文の付与率向上については、従来の「人手翻訳」を実行する以上、翻訳経費や翻訳体制の面で翻訳量に限界があり、飛躍的な向上は望めないのが現状であった。しかし、日米科学技術研究開発協力協定の中でも記述されているように、欧米と日本の間の「情報利用の相応の機会」は日本の国策でもあり、これを解決する手段としての「日英機械翻訳システムの開発」が急務となった。

1-2 JICST における開発

1982年4月～1986年3月に活動した科学技術庁プロジェクト「日英科学技術文献の速報システムに関する研究」が母体システムとなり、これに引き続き1986年4月～1990年3月に JICST 独自の開発を行った。母体システムの開発費約6億2千万円、JICST 独自の開発工数は445人月を投入し、ここで得たものは、日本語抄録を原語とした機械翻訳システムと36万語の翻訳辞書である。

2. 機械翻訳システムを活用した英文データベースの作成と提供

2-1 機械翻訳活用以前の作成方法と件数

日本情報データベース（JICST-E ファイル）の作成に着手したのは1985年1月であり、JICST オンライン情報システム（JOIS）によるデータベース・サービスを開始したのが1987年10月である。このデータベースを広く世界中に提供する目的で、1987年11月には世界規模の科学技術情報ネットワーク（STN International : Scientific and Technical Information Network）に搭載してサービスを開始した。従って、機械翻訳システムを実運用化した1991年8月までの6年8カ月の間は、「人手による翻訳作業」および「論文著者が作成した英文抄録の許諾利用」によってデータベース作成を行っていた。後者の「……英文抄録の許諾利用」とは、著者が論文執筆時に同時に作成した英文抄録を、データベース作成のために出版者や著者の許諾を得て原則として有料

で利用することで、経済的なデータベース作成法の一つであり、現在も活用している。

1985年1月から1991年8月までに、合計約117万件の英文抄録を収録したデータベースを上記の方法で作成した。この中には、1989年から入力を開始し、海外から最も望まれているグレイ・リテラチャー情報が約1万1千件含まれている。年度別・データ種別の作成件数を表1に示す。

表1 年度別・データ種別作成件数

	理工学・医学	グレイ・リテラチャー	合 計
1985年	18,454	0	18,454
1986年	162,334	0	162,334
1987年	175,701	0	175,701
1988年	186,434	0	186,434
1989年	194,718	2,062	196,780
1990年	188,966	5,163	194,129
1991年	231,182	4,299	235,481
合 計	1,157,789	11,524	1,169,313

2-2 機械翻訳を活用した英文抄録の作成

(1) システムの処理概要

システムの構成を図1に、運用処理概要を図2に示す。運用処理概要の主要項目について処理内容を紹介する。

1) 前編集作業

機械翻訳システムの実用化が遅れている原因の一つに「日本語の特性」がある。即ち、日本語には主語や目的語の省略、係り受けの曖昧さ、特殊な言い回し等がしばしば内在されているということであるが、これを目で読んで頭で判断する場合は省略や曖昧さを前後の文から推定・補完したり、思考で補うことができるため、誤解することは極めて少ない。しかし、現在の電算機ハードウェアやソフトウェアでは、人間が行う推定・補完・思考は全く不可能なため、機械翻訳システムに入力する前に、これらの不完全要素を取り除いておく必要がある。具体的には、機械翻訳システムの特性に一致するように入力文を事前に修正することである。この修正作業が「前編集作業 (Pre-editing)」である。

前編集作業を実行するには正しい日本語文法と日本語表記に関する知識があること、入力

データの専門分野知識に関する知識があること、機械翻訳システムの持つ「癖」を理解していること等が要求されるが、これらの要求を満足したとして、次に要求されることは、以下の3点である。

- 原文の意味を損なわない範囲で、可能な限りの単文化を行うこと。
- 主語や述語の欠落がなく、曖昧な係り受けを排除し、正しい句読点法に基づくしっかりした日本語構文とすること。
- 略語・略称・略号等は避けること。

このような要求を満足し得る作業者を十分確保することは不可能であり、また、生産性や作業効率をできるだけ高めるために、「日本語文章推敲機能」を組み込んでいるのが一般的で、具体的には入力文に対する形態素解析、表層チェック、語分割チェック、警告・ガイダンス表示を行いつつ、入力文をそのまま機械翻訳すると英訳に失敗するものを事前に機械的にチェックし、警告するというものである。現状から見て、この機能の発展性の余地は今後大いにある。

最終的な前編集作業は端末機を活用した日本語の Display 校正作業で、入力文であるオリジナル日本語を画面に呼び出し、上記の前編集作業結果に基づいた修正を実行して機械翻訳用日本語を完成させる。

図1 JICST 日英機械翻訳システムの構成

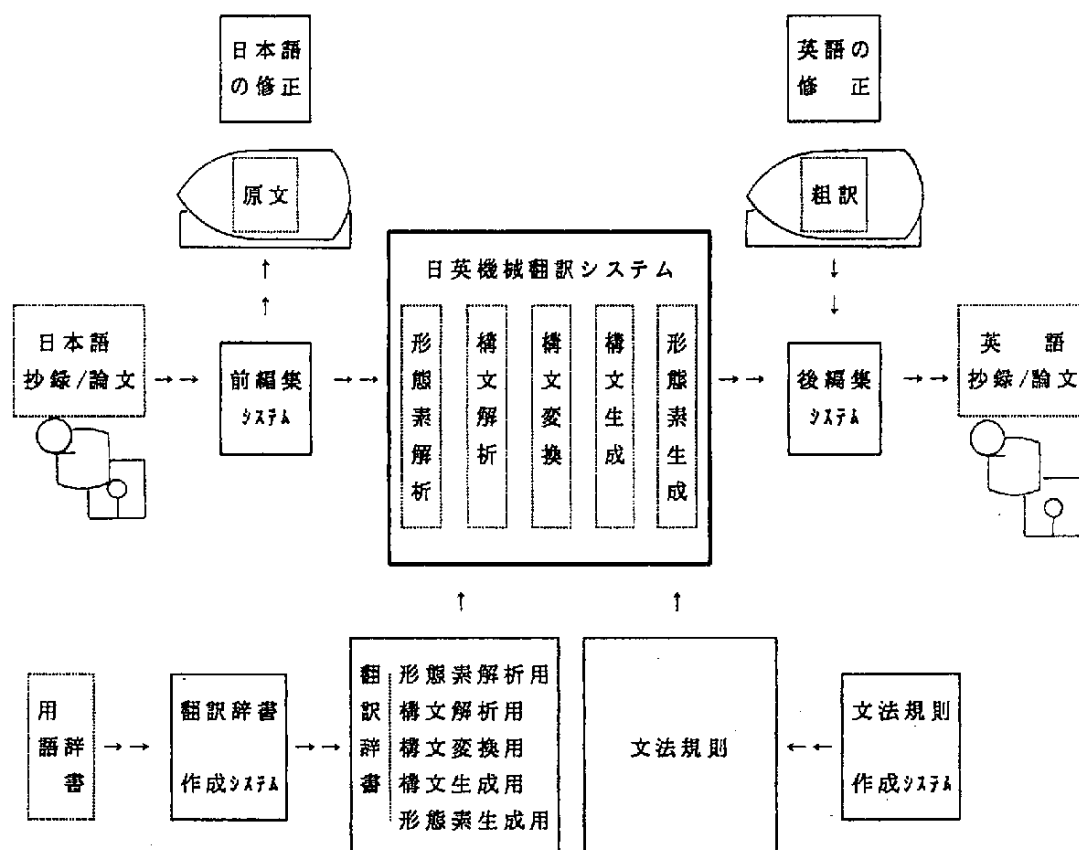


表2 翻訳辞書への品詞別登録語数 (1991.1現在)

(単位: 語)

品 詞	解析辞書	変換辞書	生成辞書
名 詞	338,037	338,037	232,227
代 名 詞	30	30	47
動 詞 (和語)	3,656	2,897	4,505
動 詞 (サ変)	9,158	9,158	-
形 容 詞	337	337	4,605
形容 動 詞	5,699	5,699	-
副 詞	413	413	1,794
連 体 詞	143	143	48
接 続 詞	81	81	34
助 詞	103	88	-
助 動 詞	111	111	9
接 語・接 辞	232	219	-
前 置 詞	-	-	158
冠 詞	-	-	3
基 数 詞	-	-	30
序 数 詞	-	-	30
単 位	-	-	5
合 計	358,000	357,213	243,495

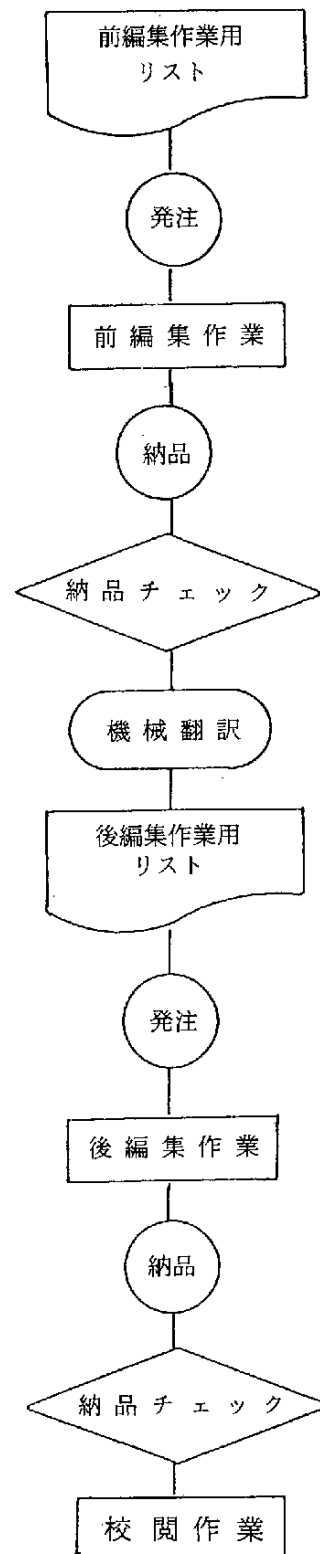


図2 日英機械翻訳システムの作業フロー

2) 機械翻訳

処理は文章単位で実行される。現在のプログラム・バージョンでの処理能力は0.27sec./語 (CPU time)、1.14sec./語 (elapsed time)、英文精度82.5%である。従って、JICSTでの機械翻訳処理は1抄録当たり、27sec. (CPU time)、112sec. (elapsed time) で実行していることになる。

3) 後編集作業

入力文が理想的に前編集され、システム・アルゴリズム、文法規則、翻訳辞書等が完璧であれば後編集作業はほとんど必要としない英文出力が期待できる。しかし、現実には上記の英文精度であるため、英文データベースとして提供できるだけの英文品質にまで、人手で校閲・修正する必要がある。この校閲・修正作業が後編集作業 (Post-editing) である。

最終的な後編集作業は端末機を活用した英文粗訳の Display 校正作業で、機械翻訳処理で得たままの粗訳を画面に呼び出し、上記の後編集作業結果に基づいた修正を実行して英文抄録を完成させる。

4) 機械翻訳用辞書の作成と維持・拡充

翻訳用辞書は、このシステムがトランスファ方式であるところから解析辞書・変換辞書・生成辞書の3種類から構成されている。解析辞書は原言語の辞書、変換辞書は原言語から目標言語へ置き換えられる単語の対や置き換え条件を与える辞書、生成辞書は目標言語の辞書という役割を持っているが、これらを作成するには膨大な時間と経費、および高度なデータ精度が要求される。因に、現在稼働中のシステムに組み込まれている36万語の翻訳辞書作成経費は、開発総経費の約半分強を占めている。辞書への登録内容を表2に示したが、3～4年以内に70～80万語程度まで増加しているであろう。

(2) システムの評価

機械翻訳システムの評価尺度として、翻訳精度、翻訳速度、運用の難易度、運用経費等がある。

翻訳精度については、目標値70%に対して82.5%という値を一応得てはいるものの、(標題のみ)については誤り率が20～40%程度であり、抄録文では「人手翻訳のための下訳として使用可」という程度である。

翻訳速度と精度とは相反するところがある。精度向上を第一の目標としているため、当初の目標値25,000語/min (CPU time) に対して1/10程度の達成率に過ぎない。従って、設計値は人手翻訳の約50%と設定していたが、現実にはほど遠いターンアラウンド時間となってしまうている。

翻訳コストの定義は幅広い。ここでは、単に「人手翻訳時の翻訳経費」対「機械翻訳時の前・後編集経費」として比較すると、前者は後者の〈4割高〉という実績がある。このことは、平成2年度経費で算定すると約1,300万円の経費節減を意味しており、上記のような翻訳精度でも大いに役だっていることが証明された。

2-3 英文データベース作成の現状

(1) 作成方式と作成件数

JICST で収集している文献の出版地を国内外に分類すると、平成2年度実績で海外文献64%、国内文献36%となっている。この36%の文献件数が約235,500件であり、平成2年度の英文データベース化量でもある。表3に作成手段別の件数を示した。

表3に示すように、抄録の付与率は37.6%と極めて悪いのが現状であり、抄録付の中で占める機械翻訳活用の件数は4.6%に過ぎない。毎年2万件ずつ付与率を増やすとしても、残る60% (14~15万件) 前後の件数を機械翻訳で処理したとしても、総てを抄録付で提供するには、あと6~7年を要するということになる。機械翻訳の処理性能の飛躍的な向上がなければこれを短縮することは不可能に近い。英文著者抄録の増加を図ることから短縮することを考える必要がある。

表3 英文データベース作成の現状

	作成方式	件数	%	件数	%	件数	%
標 題 の み	直接入力	82,454	56.1%	82,454	35.0%		
	人手翻訳	49,045	33.4%	64,536	27.4%		
	機械翻訳	15,491	10.5%				
	合 計	146,990	100.0%			146,990	62.4%
標 題 + 抄 録	直接入力	77,607	87.7%	77,607	33.0%		
	人手翻訳	506	0.6%	10,884	4.6%		
	機械翻訳	10,378	11.7%				
	合 計	88,491	100.0%			88,491	37.6%
総 合 計		235,481		235,481	100.0%	235,481	100.0%

(2) 運用体制

ここでいう運用とは上記表3の人手翻訳、機械翻訳に要する担当部署の運用のことで、直接入力についてのものは含まれていない。平成2年度の作業種類ごとの運用体制(人/月)は次のようになっている。

1) 前・後編集作業：4 機関 (委託先機関)

大手の翻訳会社2機関および学協会等団体2機関、合計4機関に委託した。それぞれの機関内での作業人員は明確ではないが、3~4人/機関程度と推定される。

2) 前・後編集 Display 校正作業：3 人 (派遣社員)

機械翻訳本体部が稼動している HITAC M-280H (14mips) に接続されている 3 台の work-station 2050/32 で作業を行っている。端末機は所内に設置されているが、電話回線さえ用意すれば全国どこでも作業することができる。作業時間帯は 9:30~17:00 で平均 22 日/月の稼動で、前・後編集込みの実作業時間は [標題のみ] で約 3 分/件、[標題 + 抄録文] で約 15~20 分/件である。

3) 納品管理・品質管理: 2 人

前・後編集作業委託先、端末機業者とのデータの受け渡し、特に、後編集作業結果の品質チェック、不良品の返品、ローテーション管理等を行っている。

4) 計算機システムの管理: 1 人

ソフトの維持・拡充、運用時の障害対策等を行っている。

5) 前・後編集作業納品物の校閲作業: 8 人

主として機械翻訳の処理対象約 26,000 件の 100% 校閲を行っている。作業時間帯は 9:30~17:00 (月~金曜日) で、年間約 16,000 文/人の作業量である。

3. 今後の課題

(1) OCR を活用した省力化と高速入力

科学技術庁プロジェクトの「日英科学技術文献の速報システムに関する研究」の第一の目的は、“日本で発生する研究論文の抄録を英語に翻訳すること”であり、1982 年から 9 年の歳月をかけて実運用が開始された。このシステムに組み込まれている翻訳アルゴリズムや翻訳辞書は、文学作品や経済学、数学、天文学等、数分野の「抄録」を除いて処理可能であり、前編集作業に時間をかければ「論文 (フルペーパー)」の機械翻訳も理論的には可能である。しかし、これを実用化する「大量情報の高速入力」には OCR が不可欠であるものの、現在市販されている日本語読み取りの OCR 性能はあまりにも悪く、誤読を修正するために多くの要員配置が必要となる。高性能の OCR 開発が待たれるところである。

(2) 機械翻訳ネットワーク・システムの開発と早急な実用化

「論文」を FAX で受信し、その機械翻訳結果を FAX で送信するというサービス・メニューが、ここ 1~2 年以内に実現するであろう。即ち、「機械翻訳サービス・ネットワーク・システム」の実用化である。平成 5 年末頃にはこのネットワークをパソコン通信でアクセスし、ユーザが入力した日本語をリアルタイムで英語に翻訳するという「道具」のプロトタイプが完成しているであろう。

(3) 翻訳コスト・提供速度/量および英文品質の大幅改善

翻訳コストの点では、人手翻訳と比較して既に 4 割程度の低下を実現させたが、当面の目標は 6 割低下であり、提供速度については人手翻訳時の半減化が目標である。翻訳コストや提供速度を大幅に改善するための大きな鍵が、機械翻訳システムから出力される英文の品質 Up であるこ

とは当然のことであるが、このための辞書拡充が急速に行われたとしても、冒頭に述べたように、日本語の特性や電算機処理能力の観点から15～20年間程度は急速な進展は望めないというのが専門家の見解である。これを突破するには、民間8社の共同で発足した「電子化辞書プロジェクト」や通産省の「第五世代コンピュータ・プロジェクト」に期待する以外に方法はないであろう。

機械翻訳システムが人手翻訳に取って代わる位に発達するのは、恐らく22世紀になってからであろうから、それまではシステムをいかに上手に使いこなすかということになる。その意味で、年間23万件の日本情報（抄録）を低コストで英文データベース化する手段はできたが、それと並行して要求されるであろう日本語論文の英訳化の手段も、高性能な日本語読み取り OCR が開発さえすれば用意されたことになる。そうすれば後はこれを上手に運用することを工夫するだけであるから、「機械翻訳システムを活用した英文データベース作成」の展望は比較的明るいと言えよう。

第2章 富士通の機械翻訳システムの開発と現状

1. 経緯

当社は英日機械翻訳システム ATLAS-I を1984年に、日英機械翻訳システム ATLAS II を1985年に汎用大型コンピュータのMシリーズ上で出荷しました。現在は WS 版の ATLAS-G も出荷しております。

これまでの当社の機械翻訳の製品および研究活動の主なものは表1のようになっています。

表1 当社の機械翻訳システムの一覧表

システムまたはプロジェクト		翻訳方式	対象言語	対象文書	時期
製品	ATLAS-I	構文解析+意味	英日	技術文書	1984年～
	ATLAS II	意味トランスラ	日英	技術文書 特に、マニュアル	1985年～
	ATLAS-G (WS版)		日英		1989年～
共同研究開発	日韓機械翻訳プロジェクト (韓国科学技術院)	ATLAS-I 方式	日韓	技術文書	1983年～ 1986年
	日独機械翻訳プロジェクト (Stuttgart 大学)	ATLAS II 方式	日独	技術文書	1983年～
展示例	EXPO' 85 多言語翻訳研究の展示	ATLAS II 方式	日本語⇒英語 仏語 独語	子供文	1985年
			英語⇒日本語 仏語 独語	世界のニュース文	

2. 利用の現状

機械翻訳システムの利用目的にはいろいろな考え方がありますが、読み方の種類を参考に分類すると以下のようにまとめられます。

①原稿作成：機械翻訳結果は一次翻訳として利用します。

後編集、ブラッシュアップ作業が必要になります。

②スキミング：翻訳結果を読んで大意をつかむために使用します。

多少、訳がおかしくてもかまいません。

③読解支援：機械翻訳の結果を参考にして読むために使用します。

辞書引きの機能も含みます。

④スキミング：大量の文書から自分の欲しい情報を探すために使用します。

日本人が使う場合には、対象文書にもよりますが、日英の場合は①が、英日の場合は②～④の目的で使用する場合があります。どの場合でも、機械翻訳システムは翻訳を支援する道具として割り切って利用する必要があります。

現在の機械翻訳システムを実用的に使うための注意事項を日英翻訳の場合を例として以下に説明します。

①対象文書の選定

工業製品のマニュアルなど明解な文章を書く必要があるものを選定します。

そのためには、マニュアルの執筆基準の検討も必要です。

システムのサポートとしては文体による文法選択機能（添付資料1）があります。

②対象文書の辞書の作成

対象文書に依存した辞書を作成します。

そのためには、用語の統一を図ることも必要です。

辞書エディタの他に多義語学習機能（添付資料2）などのサポート機能があります。

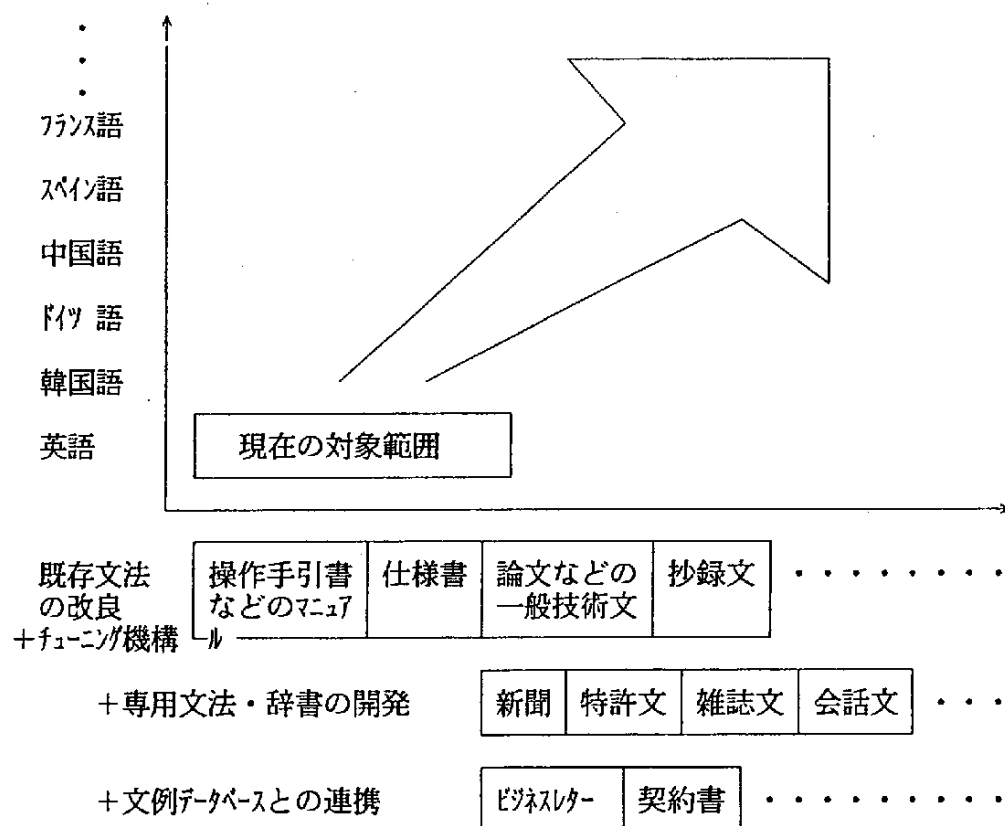
現在はまだ、ユーザに上述のような準備作業があるため、全ての対象文書にたいして導入してすぐに効果を上げるまでに至っていません。今後、文法の整備や各種チューニング機能の充実により、より容易に効果が上がるようにできると考えています。

3. 今後の機械翻訳システム

3-1 対象言語と対象文書

当社では、これまで対象言語は日英、英日であり、対象文書は製品のマニュアルを中心とした技術文書でした。今後は、以下のように、文法・辞書の整備およびユーザチューニング機能の充実、専用文法や専用辞書の整備により、対象言語と対象文書の拡大を実現していきます。

図1 対象言語、対象文書の拡がり



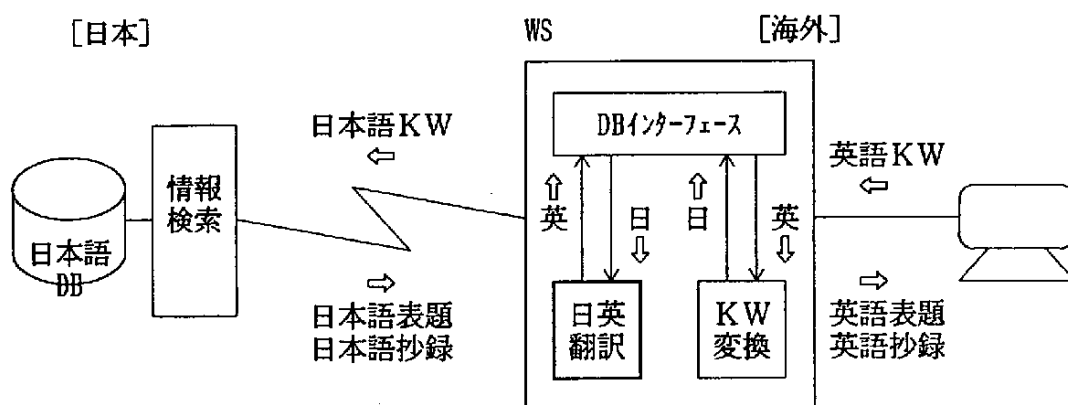
3-2 利用形態

対象とする文書や利用者などによって様々な形態で利用できるシステムが求められています。その中から、今後の機械翻訳システムの利用形態として、現在、以下のシステムを作成し、実験しています。

①オンラインDB翻訳

英語を母国語とする人たちのための日本語DBのスキャンニング・ツールとして、以下のような実験システムを作成して評価しています。

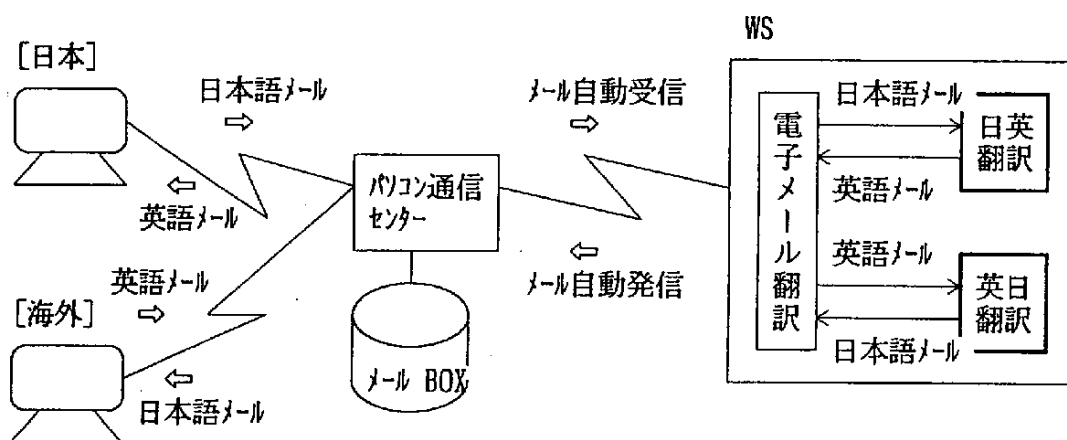
図2 オンラインDB翻訳



②電子メール翻訳

電子メールを機械翻訳センターに発信すると、機械翻訳センターが自動的に翻訳し、発信先に返信するという利用形態を実験しています。手軽に翻訳システムが利用できるという点に大きな利点があります。

図3 電子メール翻訳



4. 日本語データベースの翻訳における日英機械翻訳の課題

当社では1988年11月に COMDEX でオンライン DB 翻訳のデモンストレーションを実施し、1989年8月には IFIP の国際会議でもデモンストレーションを実施してきました。これらの経験から、日本語 DB の文書を対象にした場合、現在の日英機械翻訳について、その課題を述べます。

4-1 対象文書

日本語 DB を文章の種類により分類しその特徴を調べてみました。また、現在の機械翻訳のレベルからみた場合の翻訳可能性（スキャンニングの観点からのもの）についても評価してみました。その結果は表2のようになりました。

表2 日本語DBの機械翻訳の可能性

日本語DB	日本語文としての特徴	機械翻訳の可能性		対策
		タイトル	内容	
新聞記事	省略、並立が多い。	○	○	パターンが存在するのである程度専用文法で対応可能
雑誌記事	口語表現、会話文が多い。	○	×	口語表現、会話文の対応の検討必要
論文抄録	複雑な修飾関係や並立が多い	○	△	係り受けと並立の認識の性能向上必要
人物情報	主語省略、連用中止、単語データが多い。	—	○	特有の表現があるので専用文法で対応
企業情報	ほとんどが企業名、住所、役員名である。	—	○	単語登録が中心になる。

○：可能性あり、△：時間がかかる。 ×：すぐには難しい。

これらのことから考えると機械翻訳の対象としては、新聞記事が良いのではないかと考えられます。しかし、新聞記事も各社で特徴があるのでその点の考慮を十分する必要があります。人物情報、企業情報などは、更新頻度や一度翻訳すればあまり変更がない点からリアルタイムの機械翻訳よりサービス前に一括翻訳する時に機械翻訳を利用するのが良いと思われます。

4-2 辞書

上述の日本語DBを辞書の作成の観点から述べると、以下のような問題点があります。

①論文抄録を除き人名、地名、企業名、製品名などの固有名詞を中心とした用語の辞書の作成およびメンテナンスが大きい問題となります。人名、地名は数の問題が大きいですが、企業名と製品名は新語の確率が高く、しかも、訳語の決定が困難なものが多くあります。

②論文抄録は進歩の激しい分野の新語の登録が必要です。

③辞書の分野毎の管理だけでなく、時間での管理も必要となります。

4-3 機械翻訳の適用推進のために

今後、機械翻訳システムは、その能力を徐々に向上していき適用範囲が広がっていきます。しかし、よりスムーズに機械翻訳システムの日本語データベースの翻訳への適用を推進するためには、データベース作成者や国の機関、機械翻訳メーカー、出版社などの間で以下のような協力が必要となります。

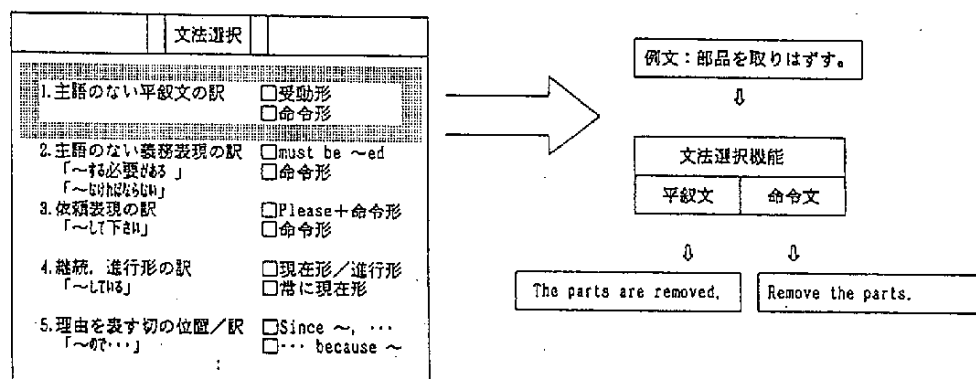
①日本語のデータベースの作成時の日本語文の作成方法を統一する。

②共同の機関で固有名詞を中心とした辞書情報の開発、維持を実施する。

【資料添付 1】

－英文のスタイルを選択できるオプション

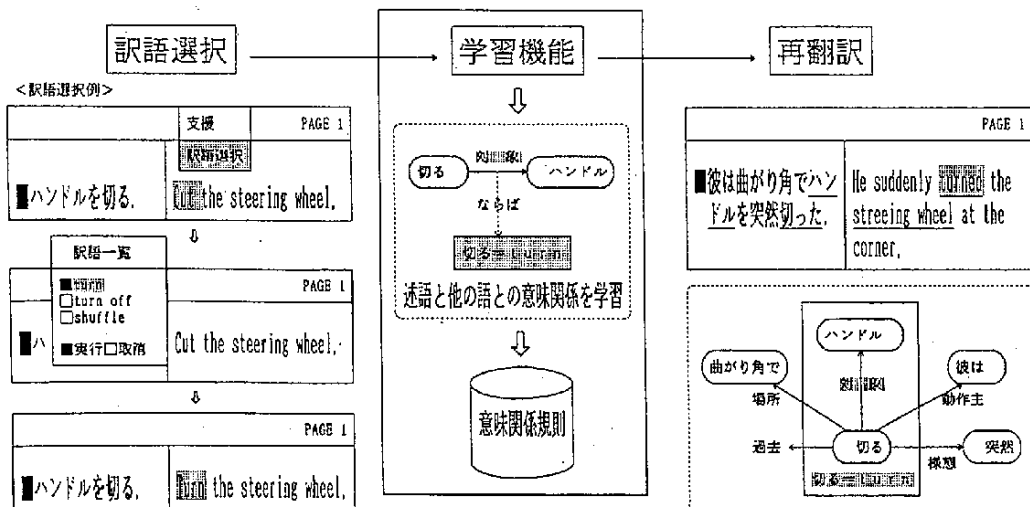
- ・ 選択項目はメーカーが提供
- ・ 翻訳文書に合わせてユーザが項目選択



【資料添付 2】

－多義語の訳し分けができるオプション

- ・ 機構はメーカーが提供
- ・ ユーザが知識を入力



第3章 シャープの機械翻訳システムの開発と現状

1. はじめに

経済、文化、政治、技術などのあらゆる分野で著しくグローバル化が進展し、コミュニケーションを促進するための機械翻訳に対する期待が高まって来ている。一方、人工知能あるいは自然言語処理の研究者・開発者は、研究開発成果の具体化として、また、企業はワープロに続く大きなビジネス商品として機械翻訳システムに対する期待を寄せている。今、使う、作る、提供する側からの要望が一致し、機械翻訳システムの実用化の機が熟して来ている。

このような背景を受け、最近、機械翻訳の開発・商品化が活発になっている。今までは、使いこなすのが難しいとの認識が一般的であったが、機械翻訳の処理技術、WSの進展、メーカーのユーザサポートの向上等により、ここ2～3年で機械翻訳に対する認識に大きな変化が現れて来ている。すなわち、機械翻訳の活用事例の発表が多くなるにつれ、使い方を工夫すれば「機械翻訳が使いものになる」という認識が定着して来た〔翻訳 1991〕。

ここでは、実用段階に入った機械翻訳システムについて、当社の英日機械翻訳システムDUE T-E/J-II〔鈴木 1991〕を中心に、開発の経緯、特徴、仕組み、効果的な利用法を概説すると共に、今後の課題について述べる。

2. 開発の状況

2-1 開発の経緯

シャープでは、自然言語処理システムの集大成として機械翻訳システムの開発に取り組んで来た。シャープにおける自然言語処理応用システムの開発は日本語ワープロに始まる。国産初のかな漢字変換方式のワードプロセッサの試作機を昭和52年春のビジネスショウで出展している。商品化の時期はライバルメーカーに先を越されたが、順調な立ち上がりを見せた。現在のパーソナルワープロの先駆けとなる、小型ワープロの開発が一段落した1983年に、機械翻訳システムの開発に着手した。日本語ワープロが形態素解析を中心にしていたのに比べ、機械翻訳システムでは構文解析、意味解析、生成技術など新たな技術課題を克服しなければならなかった。なお、シャープでは1979年にパーソナル電訳機を商品化しているが、これは、辞書的な機能が主であり、現在開発・商品化している機械翻訳システムとは別の系統のものである。

1984年には単語辞書1500語の機能モデルが完成し、簡単な英文を入力すると訳語がでるところまで進める事ができた。その後は、単語項目の増加と、内容の充実を図った。1985年には、単語数20,000語で一通りの翻訳機能を有した0次試作機が完成した。翌年には、単語数50,000語の1次試作機が完成した。当初の目標である単語数60,000語の2次試作機が完成したのは1987年の事

である。並行して、内部評価、外部評価を続け、開発着手から5年を経過した1988年DUE T-E/Jの発売に踏み切った。この時には、4種類の専門用語辞書と、OCR入力機を標準装備している。その後、ユーザの要望を反映し、半年に1回の割合でバージョンアップを繰り返した。それまで、ハードウェアは汎用のオフィス用WSを用いていたが、機械翻訳専用WSにし、単語数を66,000語に、専門用語辞書を6分野に増やして発売したのがDUE T-E/J-IIであり、1990年2月の事である。

2-2 開発のコンセプト

当社が開発を始めるころに機械翻訳の開発を行っていたのは大手のコンピュータメーカーや大学、研究所であり、ほとんど大型のコンピュータを利用していた。当社が開発に際して掲げた究極の機械翻訳システムのコンセプトは誰でもが使えるパーソナル翻訳機である。当時の技術レベルや需要を考慮し、そのステップとして以下の目標を掲げた。

- (1) オフィスで手軽につかえる小型ワークステーションによるシステム
- (2) 英語がある程度分かる人に使えるシステム
- (3) 使うほどに使い易くなる(翻訳が良くなる)システム
- (4) 1つの事業所/部門レベルで購入できるシステム
- (5) 入力から翻訳までの操作が簡単なシステム
- (6) ビジネスで使う英語文献全般に使えるシステム
- (7) 他のパソコン/ワープロなどとの接続が可能でシステムアップができるシステム

2-3 言語の選択

現在、日英、英日双方向翻訳や多言語間翻訳が一部で出て来ている。開発を始めるに当たり日英から進めるべきか英日から進めるべきかは大きな問題であった。市場調査を行った結果、需要的には英日、日英ともほぼ同じ位であり、要望的には日英翻訳への期待が大きかった。英日翻訳の場合は、粗訳でも概略を把握できるが、日英翻訳の場合は日本語の解析、出力の評価が難しいなど、使われ方や技術的な難易度を考慮して、英日機械翻訳システムの開発から進めた。日本電子工業振興協会の調査によれば 翻訳量は第1表のようになっている。

翻訳言語	割合
日本語 → 英語	47.0%
英語 → 日本語	42.8%
仏語 → 日本語	1.5%
独語 → 日本語	1.3%
日本語 → 露語	1.2%
日本語 → 中国語	1.0%

(日本電子工業振興協会)

第1表 翻訳量の調査

3. D U E T E / J - II の仕様と特徴

3-1 主な仕様

D U E T - E / J - II の主な仕様を第2表に示す。開発のコンセプトに基づいてWSを用いたシステムを実現している。また、ネットワークでの利用や移植性を考慮して、OSにはUNIXを利用している。OCRを標準装備し、入力 of 簡便性も実現している。

第2表 D U E T - E / J - II の主な仕様

項 目		仕 様
辞書	基本辞書	約66,000語(英語見出し) 分野指定可能(情報処理、電子、機械、経済、化学、時事)
	ユーザ辞書	最大約40,000語/辞書
	学習辞書	利用可能
	専門用語辞書 (オクショ)	最大約40,000語/分野(情報処理、電子、機械、 経済、科学、化学)
翻訳速度		約12,000語/時間
英文の長さ		最大70語、500文字(スペース含む)/1文
翻訳モード		対話型/バッチ翻訳、標準/大文字、である調/ですます調
ユーザ辞書登録可能品詞		動詞、形容詞、副詞、名詞、前置詞、限定詞、接続詞
本体	CPU	68030
	フロッピーディスク	3.5" (1.44MB)
	ハードディスク	80MB
OCR	フォント	117種類
	文字サイズ	6ポイント~28ポイント(2mm~10mm)
	解像度	300DPI
	認識率	99.8%(レタークオリティレベル)
適用OS		OA/UX-III

3-2 D U E T - E / J - II の主な特徴

D U E T - E / J の市場の反響を基に改良を加えたD U E T - E / J - II には以下のような特徴がある。

- (1) 使いやすい翻訳専用システム
- (2) 英文入力から翻訳文の印刷までトータルに翻訳業務をサポート
- (3) OCRの大幅な精度の改善と認識候補機能の実現

(4) 言語処理 構文解析の改良による翻訳精度の大幅な改善

(5) 1時間当たり、約12,000語の高速翻訳の実現

(6) 辞書の充実と辞書機能の充実

上の特徴を持つ機械翻訳システムをどのように実現したかについて以下に述べる。

4. 処理系

4-1 翻訳の方式

機械翻訳の方式には、入力言語を出力言語にどのように翻訳するかによって、直接方式、トランスファ方式、中間言語方式の3種類が存在する。直接方式は、辞書などを利用して、入力言語の語、句、文を直接出力言語に変換する方式のことである。上で述べたパーソナル電訳機は直接方式の代表的な商品で、メモリを消費せず、小型システムを実現できる点、また高精度の翻訳が可能な点に特徴がある。一方、融通性に難点があり、登録された語句、文だけしか翻訳が不可能である。

トランスファ方式は、入力言語を解析し中間言語を得るステップ、解析した結果を出力言語の中間表現にトランスファするステップ、その結果から出力言語の訳語を得る3つステップからなる方式で、最も一般的な翻訳方式である。解析のレベルを浅い単語のレベルから、深い文脈理解のレベルまで種々に設定でき自由度が高い。また、解析の結果を無理に抽象化しないために、きめ細かな翻訳が可能なことが特徴である。

中間言語方式は、言語に依存しない普遍的な概念構造を中間言語として設定する方法で、多言語間翻訳の実現に向いている。トランスファのステップがなく、シンプルになっているが中間言語をどのように設定するかなど課題は大きい。日本では、国家プロジェクトのC I C Cでアジア近隣諸国間の翻訳システムにこの方式が採用されている。

D U E T - E / J - IIでは、誰でも使える、きめ細かな翻訳を実現するのを究極の開発のコンセプトとしているため、トランスファ方式を採用している。また、トランスファ方式の中で、現在の技術レベルで最高の翻訳精度を得るため、意味トランスファを採用している。D U E T - E / J - IIの処理の概略フローを第1図に示す。

4-2 形態素解析

形態素処理では、語尾処理により入力された英単語を原形に戻した後に辞書を検索し、辞書に登録された情報を構文解析に引き渡す。辞書にない言葉が入力文に現れた場合は、未登録語処理により、以後の構文解析が可能なように情報を付加する。

(1) 語尾処理

入力された英文の各単語について3人称単数現在の“s”、過去形の“ed”、比較級の“er”などの語尾を取り、原形を作る。もし、複数の原形の可能性がある場合は、すべてを候補とし

て残す。

(2) 辞書検索

語尾処理で得られたすべての形態素の候補に対し、辞書の検索を行う。上の仕様で示したとおり DUE T-E/J-II では 4 種類の辞書があるが、学習辞書、ユーザ辞書、専門用語辞書、基本語辞書の順に検索する。複合語は単語より先に検索し、辞書に確定フラグが付いた項目は、個々の構成要素単語の辞書検索は行わず、複合語として確定する。

辞書の検索は、翻訳処理の基本であり、翻訳の処理速度に大きな影響があるため、各辞書にはインデックスを設けている。特に検索頻度の高い基本辞書には 2 段のインデックスを設けて、高速化を図っている。

(3) 特殊処理

今までは、通常の単語についての処理について述べた。

以下の場合、個別に特殊な処理を行っている。

1) ハイフン合成語(decision-support など)

ハイフンで結合された個々の単語に形態素処理を行う。

2) 数詞

数詞の連続を複合語と見なす。

two hundred and eighty-six : 286

3) 日付

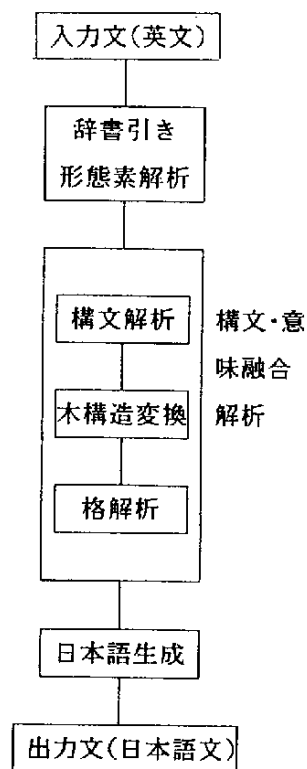
複合語と見なし、名詞と副詞の両品詞を与える。

July 1, 1980 : 1980年 7月 1日 / 名詞・副詞

4) 時間

時間のパターンを複合語と見なす。

five minutes past eight: 8時 5分



第1図 処理の概略フロー

(4) 未知語処理

4 種類の辞書を備えているが、膨大な言語現象を網羅することは不可能である。辞書検索の結果、単語が辞書に含まれない場合は、未知語として処理し、未知語に対し次の情報を付加する。

1) 接頭辞処理

接頭辞表に登録された接頭辞で始まる単語は接頭辞を除いた部分の文字列で辞書を検索し、後で接頭辞の訳と合成して訳語を生成する。辞書引きした中心語の品詞を割り当てる。

2) 接尾辞処理

接尾辞表に登録された接尾辞で終わる単語は、接尾辞を除いた文字列で辞書を検索し、後で接尾辞の訳と合成して訳語を生成する。表に記載された品詞を割り振る。

3) 算用数字処理

数名詞または数形容詞として処理する。

4) 記号を含む単語の処理

名詞として処理する。

5) 先頭が大文字の単語の処理

固有名詞として処理する。

4-3 構文解析

(1) パーザ

形態素解析で得られた入力文に対する品詞列を受け取って、句構造を出力する。パーズングの方式は、ボトムアップ・シリアル型である。ただし、高速化を図るために、失敗した規則を記憶するなどのトップダウンの予測機構を組み合わせている。

また、解析の精度を向上するために2パス方式を採用している。この方式では、文法規則や辞書の中にマイナーな文法規則、品詞、構文型の情報がある場合は、1パス目では適用せず、解析に失敗した時に実行する2パス目で適用するようにしている。このようにすることで、可能性の低い解析を無視する事なく高速解析を実現することができる。

また、動詞句、文などの構文解析のノードで随時意味解析を行うことのできる構文・意味解析の融合方式を採用している。この方式は、構文解析で出て来る不要な候補を意味解析により削除することが可能であり、構文解析と意味解析を逐次継続的に行うときに比べ、翻訳精度と処理速度の向上に寄与している。

文法規則は、文脈自由文法を基本にしているが、書き換え規則の各項にチェック関数が付加できるように拡張している。また、規則の増大を押えるためにVP/NPのようなスラッシュカテゴリを使用している。

(2) 文 法

拡張された文脈自由文法規則を約1,000個利用している。補強したチェック関数には以下のものがある。

1) 統語的非文の排除

辞書に記述された解析の型、品詞、形態属性をチェックし、非文を排除する。

2) ノードへの属性 属性値の付与

例えば、 $V \rightarrow BE + VERB$ (過去分詞)の書き換えにおいて、Vのノードに様相が受身であるという情報を付与する。

3) ヒューリスティックによる優先解釈

例えば、set the values and changes parametersの解析において、複数名詞の並列句より、後ろの並列要素を動詞として解釈する方を優先する。

4) 品詞の先読み

例えば、～asked the man to do the job の解析において、ask のノードから後方に不定詞となりうる品詞があるかどうかを先読みする。これにより無駄な解析を排除することができる。

(3) 木構造変換部

生成される構文木の数を制限するために、構文木を正規化し tree-to-tree の変換を行う。これにより、後続く意味解析を行い易くしている。木変換規則は、文脈依存文法で記述でき、規則やパターンマッチの制御が行えるようにしてある。

(4) 格解析

構文解析された結果が意味的に正しいか否かを格の照合によりチェックする。上で述べたとおり、格による意味解析と構文解析、訳語の変換を同時に行っている。格解析の処理は2つのステップに分かれており、第1ステップでは、主に訳語を絞り込むことを目的に以下の処理を行う。

- 1) 文型パターンのチェック
- 2) 格要素のチェック
 - a) 各種用法のチェック
 - b) 表層構造のチェック
 - c) 意味カテゴリー制限のチェック
 - d) 共起語のチェック

第2ステップでは、最終的な訳語を選択するために、

- 3) 深層格のチェック
- 4) 詳細意味コードのチェック

を行う。格解析に失敗した場合は、再び構文解析に戻り、他の構文候補がないか調べる。これらの格解析により、各単語の訳語、中心語の深層格、後置語、様相などが確定する。

4-4 日本語生成

格解析により決定された訳語などの情報を生成テーブルに従って生成し、並べることにより、日本語文を生成する。生成テーブルには、このための深層格と後置語の対応情報が入っている。深層格により、訳語の順序が決まり、訳語に付加する助詞なども確定する。必須格と任意格で以下のように処理が異なっている。

(1) 必須格処理

生成テーブルを参照することにより生成する。辞書に生成順序の指定がある場合はこれを優先する。深層格を持つ中心語に様相の格があれば様相の処理を行う。通常は生成テーブルに基づいて後置語を生成するが、辞書の中に後置語の指定がある場合は、辞書に記述された後置語を用い

る。

(2) 任意格処理

特殊な格を除いて、文頭の副詞句は訳文でも文頭におくなど原文の位置に準じた生成を行っている。また、副詞句、前置詞句などは非修飾語との距離が原文に近くなるように生成している。

5. 辞 書

DUE T-E/J-IIでは、総見出し約66,000語の基本辞書、基本辞書の訳語の優先順位を入れ換えた学習辞書、新しい単語などユーザが登録するユーザ辞書及び専門用語辞書の4種類の辞書を備えている。専門用語辞書は第3表に示す6種類である。

5-1 基本語辞書の内容

(1) 見出し情報

以下の情報を有している。

- 1) 見出し番号
- 2) 代表見出し
- 3) 異綴り見出し
- 4) 確定フラグ/マイナーフラグ

専門分野	語数
情報処理	22,000語
電子工学	13,457語
機械工学	18,097語
経済	30,739語
化学工学	15,545語
科学一般	26,789語

第3表 専門用語辞書

本システムで特徴的なものは、4)項で、複合語見出しがいかなる英文中においても分解不可能な場合は、確定フラグを、使用頻度の低い品詞にはマイナーフラグを与えている。これにより、無駄な解析や可能性の低い解析を低減したり、後回しにすることにより、翻訳精度の向上や速度の向上を図っている。

(2) 文法情報

以下の情報がある。

1) 品詞番号

品詞番号により複数の品詞を持つ、多品詞語の優先度を決めている。

2) 品詞名

31種類に大分類している。

3) 形態素情報

語尾変化形 形態的属性(人称、数、時制など)などの情報。

4) 品詞サブカテゴリ

品詞個別の細分類情報。

5) 解析の型

各単語の持つ構文パターン。品詞毎に必須格をパターン化している。

(3) 訳語情報

以下の情報が付加されている。

1) 訳語番号

2) 訳語及び訳語の品詞

訳語とその品詞、活用型の情報。

3) 意味属性

大局的な意味を把握するための約30種類の意味カテゴリーと、詳細な意味を決定するための約1000種類の意味コードがある。

4) 訳語文法情報

a) 用法情報

訳語を選択するための文法的条件、日本語生成のための条件など。

b) 格パターン情報

構文型のそれぞれに割り当てられた必須格に対する条件であり、以下よりなる。

イ) 必須格の表層格名(主語、名詞句、前置詞句など)

ロ) 深層格名(動作主、相手、道具など)

ハ) 生成型(後置語、語順などをパターン化)

ニ) その他の生成条件

5) 格パターン意味情報

必須格に対する制約条件(意味カテゴリー、意味コード)。

5-2 その他の辞書

ユーザに上記の情報を解放しても、混乱を招くため、一部の情報を制限している。また、専門用語は使われ方が限定しているので情報を制限している。

6. 翻訳支援機能

4～5年前に機械翻訳システムが実用化されたころには、自動翻訳システムという呼び方がはやっていて、この言葉の奥には機械から出力される訳語が人間の翻訳家と同レベルのものが期待されていた。市場に出回る機械翻訳システムの数が増加するに従って、機械翻訳システムを翻訳支援システムと位置付ける考え方が主流になって来た。特に、機械翻訳を使いこなしているユーザでは目的がはっきりしていることに加え、機械でできることを認識し、機械翻訳を支援システムとして使いこなしている姿勢が伺える。翻訳支援機能は、現在の機械翻訳の限界を打破するために、人間とシステムが協調して翻訳の作業を行うための機能である。翻訳の辞書、処理プログラムと共に翻訳の効率を決定する重要な役割を担っている。ここでは、現在の翻訳の精度と支援機能による改良の度合を見るために具体例を挙げて、支援機能について述べる。翻訳支援機能に

はプリエディット、ポストエディットの2種類があり、DUET-E/J-IIではそれぞれ9種類、4種類の機能を提供している。プリエディットでは語、句、文などのレベルでの多義性の解消が主な目的になっている。↓がエディット後の結果を示す。

6-1 プリエディット機能

(1) 品詞指定

Please note that your institution's dues ◇あなたの制度の料金が今基づく覚書
are now based on this enrollment and the を喜ばせる||この登録、及び、委員会
educational and general budged as approved によってその1990年8月会合で認め
by the comittee at its August 1990 meeting. られた教育の、そして、一般的な予算

↓

Please v_note that your institution's どうぞ、委員会によってその1990年8
dues are now based on this enrollment 月会合で認められたように、あなたの
and the educational and general budged 制度の料金がこの登録、及び、教育の、
as approved by the comittee at its August そして、一般的な予算に今基づくこと
1990 meeting. に注意してください。

(2) 文型指定

They offer patients sharper vision. それらは、患者の更に鋭い視覚を提供
する。

↓

They v4_offer patients sharper vision. それらは、更に鋭い視覚を患者に提供
する。

(3) フレーズ指定<< >>

Mr. Smith said he was not directly Smith氏は、彼が会社にあるAMレシーバ
associated with the AM receiver 設計チームと直接関係していなかった
designteam at the company and could not 、と言い、そして、レターの内容を増や
add to the content of the letters. さないであろう。

↓

Mr. Smith said <<he was not directly Smith氏は、彼が会社にあるAMレシーバ
associated with the AM receiver design 設計チームと直接関係していず、そし
team at the company and could not add て、レターの内容を増やさないのであ
to the content of the letters. う、と言った。

(4) 無翻訳指定 {{ }}

Peter Smith recongnized this fact with regard to organizations when he developed the concept of management by objectives.

◇尽きる||彼が目的による管理の概念を開発したとき、Smithは、組織に関してこの事実を認識した。

↓

Peter Smith}} recongnized this fact with regard to organizations when he developed the concept of management by objectives.

彼が目的による管理の概念を開発したとき、Peter Smithは、組織に関してこの事実を認識した。

(5) 分割翻訳指定//

Turn the nut clockwise to increase the spring force or turn it counter-clockwise to decrease the release force.

スプリング力を増大する、あるいは、リリース力を減少するために、左回りの状態に変えるために、ナットを右回りの状態に変えなさい。

↓

V3_Turn the nut clockwise to increase the spring force// or V3_turn it counter-clockwise to decrease the release force.

スプリング力を増大するためにナットを右回りに回す||あるいは、リリース力を減少するために、それを左回りに回す。

(6) 挿入指定 ()

Remarkable individual consumer spending and increasing investment on facilities and equipment by the private sector, particularly the manufacturing industry, helped make a favorable market transition.

注目すべき個々の消費者支出、増加している投資、及び、プライベートなセクターによる設備、及び、設備上の、特に製造業は、好意的なマーケット推移を行うのを助けた。

↓

Remarkable individual consumer spending and <<increasing investment on facilities and equipment by the private sector, (particularly the manufacturing industry)>>, helped make a favorable market transition.

注目すべき個々の消費者支出、及び、プライベートなセクター(特に製造業)による設備、及び、設備上の増加(して)いる投資は、好意的なマーケット推移を行うのを助けた。

(7) 削除指定(())

The weight of the collective evidence is used to determine the extent to which you can attribute specified pathological effects to a substance. ◇集合的証拠のウェイトは、範囲を決
定するために使われる||あなたが缶
詰にするものに属性は、病理学的効果
を物質に指定した。

↓

The weight of the collective evidence is used to determine the extent ((to which)) you can attribute specified pathological effects to a substance). 集合的証拠のウェイトは、範囲(あな
たは、指定された病理学的効果が物質
に起因すると考え得る)を決定するた
めに使われる。

(8) 部分翻訳指定 [[]]

While traders and economists agreed that the May producer price index figures, released yesterday, leave room for the Fed to push down interest rates, the data weren't compelling enough that the Fed would be forced to ease. ◇貿易業者、及び、エコノミストが5月
生産者物価指数が計算するそれを認
めた、と同時に||利率を押し下げるた
めに、昨日の許可連邦政府の余地を解
放した||連邦政府が弱まるように強
いられるであろう程、データは、
compellingしていなかった。

↓

[[While traders and economists agreed that the May producer price index figures, (released yesterday), leave room for the Fed to push down interest rates, the data weren't compelling enough that the Fed would be forced to ease. 貿易業者、及び、エコノミストが利率
を押し下げるために5月生産者物価指
数数字(昨日リリースされる)が連邦
政府の余地を残すということに同意
した、と同時に連邦政府が弱まるよう
に強いられるであろう程、データは、
compellingしていなかった。

(9) リライト

A high level application package that offers the capability to perform data reduction and data compression of system variables with respect to time, and store the resulting data on a bulk storage device under the control of a relational ◇時間に関してのシステム変数のデ
ータ減少、及び、データ圧縮を行う能
力を提供する高水準のアプリケーション
パッケージ、及び、ストア||関係
があるデータベースマネージャーの
コントロールの下の容積貯蔵デバイ

database manager.

ス上の生じているデータ

↓

A high level application package that offers the capability to perform data reduction and data compression of system variables with respect to time, and to store the resulting data on a bulk storage device under the control of a relational database manager.

時間に関してのシステム変数のデータ減少、及び、データの圧縮を行い、そして、関係があるデータベースマネージャのコントロールの下で容積貯蔵デバイス上の生じているデータ蓄える能力を提供する高水準のアプリケーションパッケージ

6-2 ポストエディット機能

翻訳後の内容の確認、修正のために以下のポストエディット機能を設けている。

- (1) 訳語表示選択
- (2) 次候補呼び出し
- (3) 日本語編集機能

7. 販売・教育

7-1 販 売

機械翻訳システムは現在技術的にも市場的にも成長期にあり、ユーザ、開発者、販売者が協力してより良いシステムの使い方を工夫して行く時期にある。このような時期には、販売とともにユーザの声をシステムに素早く反映することが不可欠で、DUE T-E/J-IIでは以下の対応をしている。

- (1) 定期的なソフトウェアのメンテナンス
6～8カ月に1回のバージョンアップ。
- (2) 専門用語の充実：当社独自あるいはユーザとの共同開発
- (3) 専用サービス部門の設置

DUE Tプラザと呼ぶ機械翻訳専用のサービス部門を設け、トライアル翻訳、機械翻訳のための教育、各種の問い合わせに応じている。

7-2 教 育

現在の機械翻訳システムは人間とシステムが協調して翻訳を行うことにより、翻訳作業の効率を上げることができるものである。従って機械翻訳を使いこなす人を育てるための教育が必要である。当社ではDUE Tプラザがこの役割を担っている。初級、中級、上級に分けて教育のコースが設けられており、例えば、中級コースでは次のような内容が教えられている。

(1) 他品詞語の区別方法

～ing

vg_	These are vg_running cars.	これらは、自動車を走らせている。
vd_	These are vd_running cars.	これらは、自動車を走らせることである。
a_	These are a_running cars.	これらは、走っている自動車である。
n_	These are n_running cars.	これらは、運転自動車である。

T0不定詞

tn_	I have no money tn_to go to America.	私は、アメリカに行くために、金を持たない。 →私は、アメリカに行くための金を持っていない。
tv_	Take time tv_to eat your meals.	あなたの食事を食べる時間を要しなさい。 →あなたの食事を食べるために、時間を取りなさい。

慣用句

mc_	His house is mc_as large as yours.
ms_	ms_It is no use saying that now.
mi_	mi_I see.

(2) 分割の方法

名詞句(名詞節)//文、文//名詞句(名詞節)など

That evening we came home at 6 o'clock. ◇我々が帰宅したその夜 || 6時

↓

That evening || we came home at 6 o'clock. その夜 || 我々は、6時に帰宅した。

(3) ユーザ辞書の登録方法、使用方法

9個の登録品詞 型の指定方法、4種類の自動詞、17種類他動詞、5種類の形容詞、4種類の名詞の区別、使用方法など。例えば、infuriated の場合、型は形容詞+to 不定詞で、訳語は[用1:て] 激怒するように登録する。

8. 活用状況

8-1 ユーザアンケート調査

DUE T-E/J及びDUE T-E/J-IIのユーザに1990年6月に行った利用状況に関するアンケート調査の結果を第4表に示す。

1) 業種別構成		2) 利用部門別構成比		3) 利用文書別構成比	
業種名	構成比	利用部門名	構成比	利用文書種類	構成比
製造	32.4%	技術	22.7%	マニュアル	21.0%
商業・流通	11.4%	企画	16.2%	技術文献	18.3%
ソフトウェアハウス	9.7%	外国	7.6%	雑誌	11.4%
官公庁・学校	8.1%	製造	5.9%	海外レター	11.0%
運輸・通信	5.9%	翻訳	5.4%	海外レポート	10.0%
翻訳業	5.9%	社長室	5.4%	カタログ	8.2%
不動産	5.4%	営業	3.8%	その他	20.1%
金融	4.3%	その他	33.0%	合計	100.0%
その他	16.9%	合計	100.0%		
合計	100.0%				

第4表 DUE T-E/Jの利用状況（回答185社）

8-2 有効活用事例

上記ユーザの中で、機械翻訳を有効に活用しているユーザの事例を紹介し、それらの共通点を整理する。

(1) 翻訳会社

1) 導入のきっかけ

仕事の効率化・スピードアップを図りたい。

大型の汎用コンピュータに比べ、コスト、使い勝手が良い。

2) 利用方法

一連の翻訳作業の流れを分業化して利用。

(OCR入力 → 翻訳 → 編集)

3) 主な導入のメリット

a) 人件費や外注費等のコストが大幅にダウン。

b) 翻訳作業のスピードアップ（約6倍）と経営効率のアップ。

c) 訳語・品質の統一。

(2) 印刷会社

1) 導入のきっかけ

大量の英文を短時間で翻訳／印刷したい → 翻訳の機械化。

2) 利用方法

翻訳結果のエディット（編集）の他はすべて自動化。

顧客毎にユーザ辞書を作成。

3) 主な導入のメリット

a) トータルでコストパフォーマンスの高い生産システムを実現。

b) 用語や文体・訳調など品質の統一。

c) 顧客毎にきめ細かい翻訳が可能。

(3) 自動車会社

1) 導入のきっかけ

海外情報（技術文献など）を即座に翻訳したい。

2) 利用方法

OCRで入力し、粗訳を関係部門へ回覧。

ユーザ辞書を作成、翻訳精度を徐々にアップ。

難解な英文はエディット（編集）を加える。

3) 主な導入のメリット

a) 外注費の削減

b) 翻訳量・スピードのアップ → 情報を素早くつかめる。

c) 対訳で回覧→理解度がアップ。

(4) 大使館

1) 導入のきっかけ

本国からの英文による情報の増加。

2) 利用方法

OCR、フロッピーディスクによる入力。

DUE Tで大量に粗訳にかける → 編集し完全な訳に。

パソコンのDTPソフトで仕上げ → 印刷。

3) 主な導入のメリット

a) 外注費の削減

b) MS-DOSファイルの英文をダイレクトに翻訳→大幅な時間短縮。

以上の機械翻訳導入に成功したユーザに共通することは以下の6点であり 今後機械翻訳の導入を考えているところは 導入に際してこれらの各点を十分に検討をしておく必要があろう。

- (1)目的が明確である
- (2)良き理解者/強力な推進者がいる
- (3)専任者がいる
- (4)機械化へのポテンシャルが高い→システムの運用
- (5)入出力の効率化
- (6)分かり易い導入メリット

9 D U E T - J / E

シャープではD U E T - E / Jの開発が一段落したところで、日英機械翻訳システムD U E T - J / Eの開発に乗り出した。平成2年10月17日に新聞発表を行い、平成3年秋までに商品化の予定である。主な仕様を第5表に示す。

D U E T - J / Eは基本的な操作性、ハードウェアは英日翻訳システムと共通になっている。ただし、日本語は省略が多く、語順の制限が英語に比べ緩いという特徴に合わせるため、パーザはボトムアップの平行方式を採用している。意味トランスファ型で構文・意味融合型であるところは英日翻訳と共通である。

辞書英日と同様4種類を備え基本語辞書の見出し数は7万語を用意している。日本語の場合、漢字仮名交じり文を正確に形態素解析することが、第一条件になるため、長単位語などを充実させている。

また、日英翻訳の利用場面を考えた場合、相当に普及した日本ワープロで入力した文章を翻訳機にかけることが多いとの判断から、当社のワープロ書院で作成した文書が直接翻訳機で読めるような

インタフェースを付加している。OCRは英日翻訳の場合と異なりオプションになっている。翻訳支援機能を充実させ、翻訳の効率の向上を図っている点は英日翻訳の場合と同様である。ただ

項 目		仕 様 内 容
翻訳方式		意味トランスファー方式
翻訳モード		一括/対話型翻訳
辞書	基本語辞書	日本語見出しで約70,000語
	学習辞書	利用可能
	ユーザ辞書	最大40,000語/辞書
	専門用語辞書 (オプション)	情報処理、電子工学、 機械工学、経済
翻訳速度		5,000語/時間
翻訳支援機能		左右対訳表示、日英対応語句 表示、訳語表示選択、語義指 定、翻訳文型指定、部分翻訳 指定、無翻訳指定、係り受け 表示、係り受け変更、単語検 索、文書校正
CPU		MC68030
メモリ		16MB
ディスク		167MB
OCR (オプション)	認識文字	マルチフォント
	認識対象	約4,000字
	文字サイズ	6~28ポイント
	認識率	99%

第5表 D U E T - J / Eの仕様

し、省略語の確認機能、冠詞の設定機能など日英翻訳独自の機能を付加している。日本語の入力ミスやブラッシュアップのための校正機能も英日翻訳にはなかった機能である。

10. 今後の課題

シャープでは、英日機械翻訳システムが世に好評裡に受け入れられ、日英機械翻訳の発売に近い時期に実現すべく準備を続けている。多数のユーザから上がってくる要望を集約すると次の8つになる。これらは、いずれも今後の課題として取り組んでいるものである。

- | | |
|-----------------------|-------------------|
| (1) 機械翻訳の精度の向上 | (5) 日英、英日双方向翻訳の実現 |
| (2) ユーザインタフェースの向上 | (6) 他社マシンへのソフトの移植 |
| (3) ユーザ辞書、翻訳ノウハウの相互利用 | (7) 小型化、パーソナル化 |
| (4) 専門用語辞書の充実 | (8) 低価格化 |

(1) では、多義性の解消をいかに行うかがポイントである。他品詞語、並列句、「の」を含む名詞句などの処理を適確に行う必要がある。照応、省略など文章理解を問題にする前に、文脈情報を利用してこれらの多義性の解消を図るべきであろう。また、長文の処理も多くが多義性が出現し、処理時間もかかることから、ユーザインタフェースとの問題も極めて重要な課題である。これらに対しては、現在も各種の方法が試みられている。例えば、分野やテーマを推測または設定し、訳語を選択したり、ヒューリスティックな優先規則を適用したり、例文ベースを利用したり、文書のレイアウト情報を利用したり、定理証明の推論を用いたりするなど。

(2) では、プリエディット、ポストエディットを更に使い易くするようなインタフェースが望まれている。例えば、多義性を絞り切れない場合、区別し易いような情報を付加しながらユーザに問い合わせる、音声などのマルチメディアを使うなど。また、いったん修正実行した翻訳ノウハウをシステムが自動学習し、次回からの翻訳に行かせるようにするなどの、学習機能が課題となる。

(3) は辞書やノウハウの普及、流通に関するものである。著作権などを含め、今後の課題である。公的な機関として、4月に発足した日本機械翻訳協会がこのような便宜をはかることを設立の目的の一つにしている。

(4) はユーザが増加するに従って、現在提供している英日翻訳6種類の専門用語以外の辞書の要望が増えていることによる。今後、更にユーザが増加するに従い、課題となる。

(5)、(6)、(7)、(8) は上記の4つの課題が解決されれば、ユーザが更に増え、ハードウェアやWS、インフラストラクチャの充実、向上により解決される問題であると考えている。

機械翻訳を通じ、情報の交流を促進し国際親善に寄与することを望む。

【参考文献】

- [翻訳 1991]、"特集 機械翻訳のいま" 翻訳の世界 バベル・プレス pp25~72 N0.2 (1991)
[鈴木 1991]、"機械翻訳システムの現状と将来" 第40回システム制御情報講習会 pp63~74(1991)

第4章 RSC & CLIにおける機械翻訳システムの現状

1. システムの開発経緯

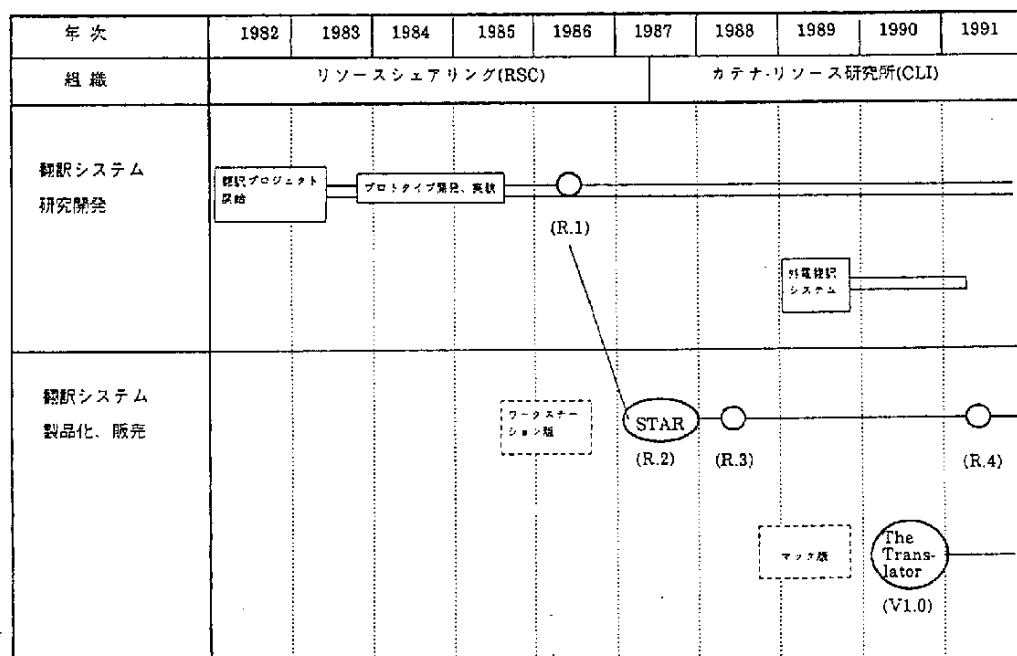
1980年代初めから、リソースシェアリング機(RSC)では日本語を中心とする自然言語処理の研究および実験システムの開発に着手し、言語処理技術の蓄積、発展に力を注いできた。こうした中で、1982年頃から機械翻訳プロジェクトがスタートし、英日翻訳システムのプロトタイプ開発、実験が3～4年に渡り行われ、本格的な製品化に着手した。

その後、1987年に翻訳プロジェクトの研究開発、製品化組織が、新たに設立された㈱カテナ・リソース研究所(CLI)に引き継がれ、英日翻訳システム「STAR」という名称のもとに、一般ユーザに製品パッケージとして本格的な販売が開始された。発売当初のバージョンは、リリースー2(R.2)であったが、その後何度かのバージョンアップを行い、まもなく大幅な変更が加えられたリリースー4(R.4)を提供する予定である。

また、「STAR」で用いられている翻訳エンジンをもとに、1990年夏アップル社の Macintosh 用の英日翻訳システム「The Translator」を㈱アバクスとの共同開発により製品化した。

「The Translator」は、Mac がもつ優れたユーザインタフェースを最大限利用して、簡便で判り易い翻訳環境を実現している。そして、翻訳システムの利用者層の拡大を図るべく、パーソナル・ユースでのニーズに焦点をあわせ、本格的な翻訳システムとしては低価格のパッケージソフトとして販売を行っている。

図1 翻訳システムの開発経緯 (RSC&CLI)



2. システムの概要

2-1 WS 版翻訳システム「STAR」

英日翻訳システム「STAR」は、RCS および CLI において永年培ってきた自然言語処理技術に関する研究開発成果を、一般ユーザの利用に供すべく英日翻訳システムとして製品化したものである。

機械翻訳システムは、自然言語処理技術を応用したアプリケーションとして最も早く実用化されつつある。しかし、人間や社会にかかわる万象、社会的・個人的歴史などの背景を持つた自然言語を全てにわたってコンピュータで扱うことは、本質的に困難である。翻訳システムにおいても、複雑な処理機構の全てを利用者に対して閉ざしたまま、完全な翻訳結果を提供することは、現在の技術的・理論的水準を超えたレベルでも相当に無理がある。

このため、「STAR」の開発・製品化に際し、我々はシステムの骨格として、理論的・技術的に確実なものを採用し、利用者が様々なレベルでシステムに手を加えることを最大限保証することにより、システムの活用面の可能性を幅広く追求できることを中心的なねらいとしている。

したがって、「STAR」は現時点での可能な技術レベルとして最大限の翻訳精度を保証しているが、そのみでは自然言語を扱うには未だ不十分であるという限界を乗り越えるためにも、システムの開かれた可能性を利用者に提供していることが、「STAR」の最大限の特徴といえる。

現在、UNIX オペレーティングシステムを使用した各種ワークステーション用のソフトウェア・パッケージとして提供を行っている。

また、「STAR」は、製品として販売後も翻訳精度、性能、操作性向上を図るため何度かのバージョンアップを積み重ねてきており、現在リリースー3 からリリースー4 へと移行中である。

(1) 特 徴

「STAR」は、高度な自動翻訳機能を核として、翻訳精度を高めるための豊富な翻訳サポート機能、様々なオペレーションに対応するための柔軟なユーザインタフェースを備えている。また、ユーザの利用形態に応じて拡張、変更することが容易であり、最適な翻訳処理システムの構築を保証するオープンエンドなシステムである。

1) 高速の翻訳スピード (RISC マシン……10万語/時)

一般的なマニュアル、新聞、雑誌記事などの文章を一時間に10万～20万語以上のスピードで翻訳。

2) 大量文書の翻訳も一括して処理可能

一文単位に翻訳を行い、翻訳結果を直ちに画面に表示して確認したり、ファイルに出力。

3) マルチユーザによる共同利用に最適

複数のワークステーション、あるいは端末から同時に複数のユーザが利用可能。辞書の共有、個別の使い分けが容易。

4) 辞書を全面的に開放し、ユーザ独自の辞書修正・追加が可能

システム辞書、オプション辞書の開放を図り、翻訳の質、利用分野の専門語の拡充のために、辞書の変更、追加が自由。

5) 翻訳ニーズに合わせたシステムのカスタマイズ化が容易

本システムは、コマンドをベースとしており、UNIX の豊富なコマンドと組み合わせて、ユーザの希望する翻訳処理に合わせたシステムの構築が容易。

6) 翻訳方式は高度で確実なアルゴリズム

統語論的な言語解析技術をベースとして、意味論的な格支配構造なども取り込み高度でありながら確実なアルゴリズムを実現。

(2) 主な機能

「STAR」は、リリースー3 からリリースー4 へのバージョンアップに伴い、機能の改良・追加が行われているが、リリースー3 の機能としては、次のようなものをユーザに提供している。

1) 標準的な翻訳処理を行うための会話型翻訳モニタによるメニュー形式での操作環境

2) テキスト編集、翻訳結果の解析、辞書参照等をサポートするマルチウインドウ形式の後編集機能

3) システム辞書の全内容について変更、追加が可能なユーザ辞書メンテナンス機能

4) 複数の専門語辞書などを同時に利用するためのオプション辞書メンテナンス機能

5) 翻訳、辞書ユーティリティ、各種ツールとして、表1に示すようなコマンドの直接実行機能

表1 ユーザ・コマンド一覧

「翻訳」

No.	コマンド	機能	備考
1	star	翻訳モニタ	
2	translate	翻訳コマンド	
3	generate	生成コマンド	
4	post	後編集	
5	snart	翻訳フィルタ	
6	starvi	翻訳テキスト作成用vi	

「辞書ユーティリティ」

No.	コマンド	機能	備考
1	cudic	ユーザ辞書のコンパイル	
2	codic	オプション辞書のコンパイル	
3	lookey	ユーザ、オプション辞書内容の表示	対話形式
4	xlook	ユーザ、オプション辞書内容の表示	バッチ形式
5	edict	辞書修正用vi	
6	verdict	ユーザ辞書ソースのエラーチェック	
7	sortdict	ユーザ辞書ソースのソート	
8	mergedict	ユーザ辞書ソースのマージ	
9	diffdict	ユーザ辞書ソースの比較	
10	dhelpt	ユーザ辞書記述の説明	
11	slookup	辞書ソースの表示	
12	usr2opt	ユーザ辞書からオプション辞書へのソース変換	

「各種ツール」

No.	コマンド	機能	備考
1	chktext	入力テキストのチェック	
2	digest	翻訳テキストからの特定辞書の作成	
3	unreg	未登録語の表示	
4	mspell	vi中からの綴りチェック	
5	starhelp	翻訳システムのヘルプ・ユーティリティ	
6	starman	コマンド・マニュアルの表示	manスター版
7	starenv	翻訳環境の表示	envスター版
8	starap	キーワードによるコマンドの表示	
9	starclean	翻訳中間ファイルの削除	

(3) システムの実行環境

「STAR」は、メーカーから数多く製品化されている UNIX ベースの主要なワークステーションで利用可能である。「STAR」を使用するための環境条件は、次のとおりである。

【ハードウェア】

- | | |
|----------|--|
| ●本 体 | UNIX系OSのワークステーション、ミニコンピュータ
メインメモリ 4MB以上 |
| ●ハードディスク | AP使用領域 10MB以上 |
| ●端 末 | 日本語ディスプレイ |

【ソフトウェア】

- | | |
|----------|--------------------------|
| ●OS | UNIX系OS(SYSTEM-V、4.3BSD) |
| ●日本語処理機能 | 日本語フロントエンドプロセッサ |

【代表システム例】

既にインストール済みの主要なワークステーションとして、以下のものがある。

- | | |
|----------------|----------------------------|
| ●ソニー | NEWSシリーズ |
| ●NEC | EWS4800シリーズ |
| ●日立 | 2050シリーズ |
| ●住友電工ワークステーション | USTATIONEシリーズ |
| ●オムロン | LUNAシリーズ |
| ●松下コンピュータシステム | BEシリーズ |
| ●東芝 | AS-3000、4000シリーズ |
| ●日本サンマイクロシステムズ | SUN-3、4、SPARC station シリーズ |
| ●DEC | VAXシリーズ (OS:ULTRIX) |

(4) システムの仕様

「STAR」の基本的な仕様は、次のとおりである。

- | | |
|-----------|--|
| ●翻訳方式 | 構文トランスファ方式 |
| ●翻訳速度 | (CISC マシン) 1～2万語/時
(RISC マシン) 5～10万語/時 |
| ●単語辞書 | システム辞書(一般語) 4万語
オプション辞書(専門語)
コンピュータ分野等 |
| ●文法規則 | 3,000ルール |
| ●システム記述言語 | C |
| ●システムサイズ | 7MB |

2-2 Mac 版翻訳システム「The Translator」

Mac 版の英日翻訳システム「The Translator」は、ワークステーション用に数多くの実績を持つ「STAR」の翻訳エンジンをベースに、Mac の優れたユーザインタフェースを利用して、簡単な操作で手軽に翻訳が行えるシステムである。

特に翻訳結果の校正、編集機能として、訳語置換、別解選択、部分翻訳などをサポートし、翻訳精度を高めるために辞書登録、訳語の学習機能などを数多く組み込んでいる。

また、Mac 用には、サードパーティーからコストパフォーマンスに優れた OCR ソフト（文字読み取りソフト）が数種類発売されており、この OCR ソフトと「The Translator」の組み合わせにより、ドキュメントの翻訳が手軽に行える。

(1) 主な機能と特徴

「The Translator」は、次のような数多くの特徴を持っている。

- 一文単位の翻訳により、編集に便利。

翻訳したい英文を一文単位で適切な日本語に翻訳。したがって、レポートや資料作成などの文書構成、編集に便利。

- 複数の解釈にも対応。

英文の言いまわしや使われる単語によって複数の解釈ができる場合、複数の訳中から適切な解釈を選択可。

- 新しい訳語の置き換えも自在。

英文と日本語文の単語の対応関係をみたり、訳語を他の語や新しい語に置き換えることが簡単。

また、置き換えた訳語を次から最優先で選択する学習機能。

- 辞書の追加が簡単。

新しい単語の登録が簡単。登録された単語はすぐに翻訳結果に反映。また、辞書にない単語が翻訳したい英文に含まれていても、その単語の品詞を推定して日本語文を生成。

- 必要なものは原語のまま翻訳。

英文の中で、そのまま残したい名詞句（原著名、出典名等）など、一部分を原語のまま生かした日本語文を作成。

- 再翻訳や日本語文の編集も容易。

画面上で、英文テキストを編集・訂正して再翻訳させたり、翻訳結果の日本語文を直接編集可能。

- 複数の翻訳を同時に処理。

複数の英文テキストをオープンし、並行して翻訳。

- マルチファインダの環境で実行可能。

マルチファインダのもとで、バックグラウンドで翻訳を実行したり、別のアプリケーション（ワープロなど）と切り換え、翻訳。

- 翻訳結果の保存形式は2通り。

翻訳結果の保存には、ワープロ、DTPプログラムなど利用可能なテキストファイル形式と、翻訳処理を再度実行するための翻訳ファイル形式の2通りが用意。

テキストファイル形式では、英文と日本語文の対訳形式、英文のみ、日本語文のみの3種類。

- プリントアウトの形式も選択可。

翻訳結果のプリントアウトも、英文と対訳日本語文の対訳形式、また日本語文のみの2種類が選択。

- 各種テキストファイル、OCRに対応。

印刷物、タイプやワープロで打たれた英文などを直接、OCRシステムで読み込み翻訳。Macのディスクアクセサリ（スクラップブックなど）に取り込まれた英文テキストや、もちろん、キーボードで入力された英文の翻訳も可能。（OCRシステムはオプション）

（2）システムの実行環境

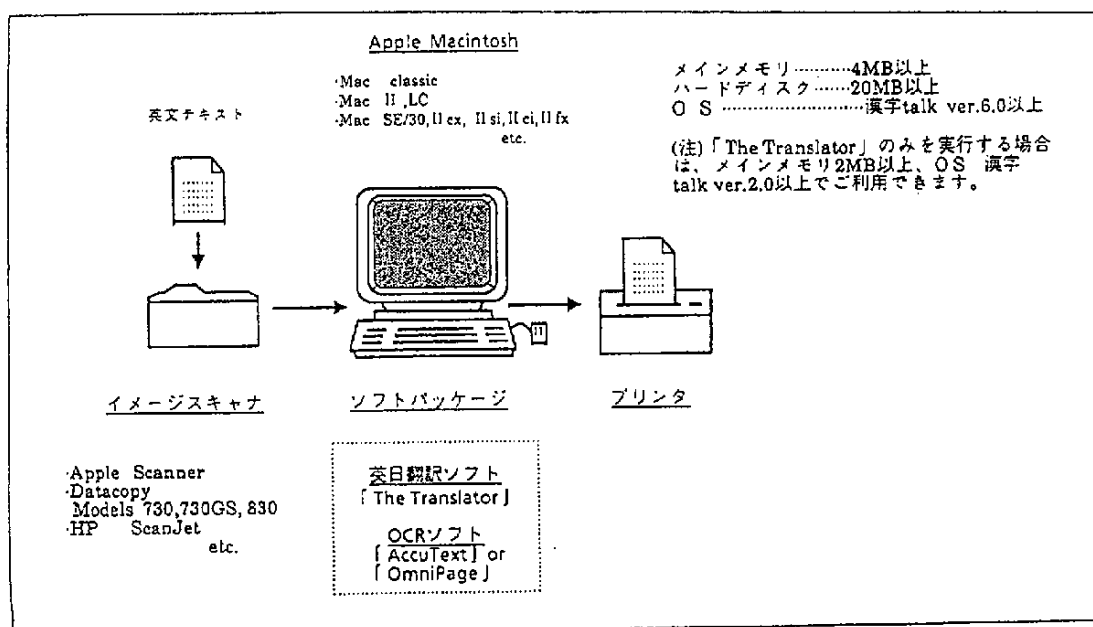
【ハードウェア】

- Apple Macintosh シリーズ
- メインメモリ 2MB以上（ユーザーズメモリー）
- ハードディスク 10MB以上（AP使用領域）

【ソフトウェア】

- OS 漢字Talk (Version 2.0以上)

図2 Mac用英日翻訳システムの全体構成図



(3) システム仕様

● 翻訳方式	構文トランスファ方式
● 翻訳速度	1～2万語/時
● 単語辞書	システム辞書 2.5万語 学習辞書(学習機能により逐次登録) オプション辞書……コンピュータ分野等
● 文法	3,000ルール

3. システムの利用状況

当社の翻訳システム「STAR」および「The Translator」は、表2に示すような特徴の違いがあり、それぞれの特徴を生かした利用が行われている。

3-1 「STAR」

「STAR」のユーザは、現在電機メーカーを中心とする製造業が半数を占めてをり、この他大学、公共機関、民間の研究所、翻訳会社、情報サービス会社などで使われている。

対象分野として、一般文書からテクニカルドキュメントまで様々な分野の翻訳に利用されているが、比較的マニュアルに関するものが多いようである。

NHKでは、衛星放送の海外ニュースの日本語テロップ作成に「STAR」を導入した翻訳作業を試験的に進めている。

また、研究実験段階であるが、外電ニュースの翻訳に「STAR」を用いて英文ニュースを受信すると同時に日本語への翻訳を行なう即時翻訳システムを実現しつつある。このシステムは、NHKとCLIの共同研究プロジェクトとして行なわれており、この研究成果は実用システムとして近い将来、翻訳システムの特化したアプリケーションの一例として公開される予定である。

3-2 「The Translator」

「The Translator」は、低価格であるため比較的手軽に購入され、パーソナル・ユース向けの翻訳ツールとして利用されている。

本製品については、販売元であるカテナ㈱において、定期的に講習会が開かれており、一時間程度の講習で一通りの操作が理解できるものとなっている。

利用者は、まもなく1,000ユーザ程度になりつつあるが、一般オフィスでの文書の翻訳、個人レベルでの少量の文書翻訳などに多く使われている。

また、一部の翻訳者においては数千語の辞書登録を行い、前編集を絡め相当深いところまで使い込むことが行われている。ただし、「The Translator」は、ユーザ辞書(基本となるシステム辞書を変更・追加できるもの)を開放していないため、辞書登録による翻訳精度アップには、多少限界がある。なお、こうした「The Translator」のパワーユーザに対しては、「STAR」への移行

ができるよう作成した辞書の変換ツールなどを整備する予定である。

表2 WS版「STAR」とMAC版「The Translator」の比較表

	WS版「STAR」	MAC版「The Translator」
翻訳スピード	10~20万語/時間	1~2万語/時間
入力テキストのサイズ	入力テキストファイルの大きさの制限は、特になし	入力テキストファイルとして扱える大きさは、200KB程度まで(但し、AP実行メモリサイズを拡げることにより1MB程度まで可能)
ユーザ数	複数ユーザによる同時使用が可能 また、複数の翻訳処理を並行して行うことも可能	シングル・ユーザによる利用のみ バックグラウンドでの翻訳は可能
操作性	コマンド形式により、非定型処理にも対応可能(様々な入力テキストや出力指定など)	メニュー方式による定型的な処理 簡便で使いやすい
辞書	ユーザ辞書、オプション辞書を開放 (ユーザ辞書は、利用分野に合わせて翻訳精度を高めるために、基本辞書であるシステム辞書の記述を変更・追加するもの)	オプション辞書のみ開放
スペルチェック	テキストエディタ内に組込み。	他のAP、DAに組込まれたスペルチェック機能を利用
エディタ	UNIXのjviコマンドを拡張したものを使用	簡単なエディタ機能が組込まれている。
システムのカスタマイズ化	ユーザの使い方、翻訳処理に合わせてシステムのユーザインタフェースを変更・追加したり、辞書をチューンアップすることが可能	不可
利用目的	ビジネスユーザ向 (大量翻訳、即時翻訳に最適)	パーソナルユース向 (個人で手軽に使用)

4. 今後の課題

翻訳システムは、今後ますますその需要が高まるものと予想されるが、現在の翻訳精度は、まだユーザにとって見れば不十分でありさらにレベルアップするすることが必要である。

このため、当社でも主として以下のような作業を通じて翻訳システムの精度向上、操作性の改良に努めている。

● 翻訳システムの処理プログラムの改良

翻訳プロセスは、大きく解析、移行、生成の3つのフェーズからなるが、このうちの生成（日本語文を作り出すプロセス）処理の改良に力をいれ、できるだけ判りやすい日本語文を出力できるようにする。

● システム辞書の拡充

基本辞書となるシステム辞書の量的拡大（エントリー数 2万語→5万語）、質的拡充（品詞、訳語のウエイト付けおよび、意味属性の付与による訳し分け）を推進

● オプション辞書のリリース

専門語を分野別に網羅したオプション辞書をニーズの高い分野から順次提供していく予定（医学、薬学、化学、機械、経済、など）

● ユーザインタフェースの改良

翻訳時の前編集、後編集および辞書登録・修正などに関する操作を簡単で使いやすいものにすべく、さらに改良を加える。

● 翻訳の前処理ツールの整備

翻訳の対象となる文書には、様々なものが考えられるが、翻訳を実行する前に翻訳時に想定しない処理、意味のない処理などを行わないように、あらかじめテキストのフォーマット、センテンスをチェックし、必要に応じて前処理ができるようにする。

（文章の整形、スペルチェック、翻訳対象範囲／対象外の指定、文末処理、文種指定など）

第3部 機械翻訳システムの周辺の現状

第1章 学術用語集のCD-ROM化

第2章 シソーラスの開発と利用

第1章 学術用語集のCD-ROM化

1. はじめに

学術用語集とは文部省が編纂している学問分野別の専門用語の日英対応表であり、昭和29年発行の数学編以来、28の分野で作成されている。

学術用語集そのものは専門用語集であり、学生や研究者にとって必要なものであるが、一方では初等中等教育で使用する術語を定める、という性格も持っている。そのため、

(1) 各学問分野での研究の進展に伴う改定

(2) 分野間での用語の調整

ということが重要になってくる。学術用語集は学協会の主導的な編集作業により、適宜改定が加えられる。

この際に編集委員は他分野で同じ英語を用いていないかどうか、等の調整をしながら、新たな用語の制定を行っていくことになる。

本稿で紹介するCD-ROMは、主として用語を制定しようとする人が座右に置いて利用するための道具として、試作したものである。

2. 学術用語CD-ROM検索システム——設計の考え方

学術用語データベースの制定支援の中でエンドユーザに対して検索機能を提供するサブシステムがSci-Terms/CD-ROMである。Sci-Terms/CD-ROMは、用語制定作業を行なう学会員等が手近なパーソナルコンピュータで簡単に検索できることを第一の目的としているから、そのハードウェア構成は必然的に現在の我国の研究者に最も普及している機種を選択することになった。すなわち、NEC製のPC-9800シリーズパーソナルコンピュータを基本機種として設定した。しかし、研究の内容としてはパーソナルコンピュータの機種にとらわれず、今後なるべく多様な機種に適応して行けるよう配慮している。また、制定作業中に要求される検索機能の全容については完全には解析されていないので、設計段階ではおよその要求を推察しながら設計を行なわざるを得なかった。ここで特に留意した点は、CD-ROM中に納めるデータについては、一度プレスしてしまえばその後の修正変更はなかなか難しいが、一方、それを活用する検索ソフトウェアは後で修正が可能なことである。そこでまずCD-ROM中のデータについては種々の検討を行ない、検索上で将来でてくるかも知れない要求に対処できるものになるよう配慮した。パーソナルコンピュータ上の検索ソフトウェアについては、後述のように現時点では様々な制約があるので、ある程度妥協せざるを得ない面があった。なお、同一のCD-ROMに対して検索ソフトウェアをより一層洗練させていくことは、本研究の継続作業として行ない得るものである。

3. CD-ROM の現状

現在種々市販されている CD-ROM では、3 段階程度の索引を使用して検索することが行なわれているようである。これはちょうど昔の索引順編成 (Indexed Sequential) ファイルで、3 段のインデックスを作った場合の内部形式と似たような形である。検索を高速化する上で、このすべてのインデックスファイルをメモリー上あるいはハードディスク上に持って行きたいところであるが、PC-9800 のアーキテクチャではそれを強制することはできない。特に主記憶の最大値は 640KB であり、しかも仮想記憶がなく、インデックス全体の主記憶常駐は不可能である。そこで第 1 段階目のインデックス (主インデックス) のみ主記憶常駐とし、2 段階目以降のものは各ユーザのハードウェア構成に応じて適宜メモリー常駐 (RAM DISK を含む)、ハードディスク常駐あるいは CD-ROM の直接アクセスを使い分けることにした。

また CD-ROM はアクセス速度が遅いこととプログラムの規模の問題から、検索条件を分解して個別の検索集合を作りその論理演算で最終的な検索結果を得るという、大型コンピュータの情報検索システムにおける常套手段を採用しにくい。インデックス検索のために何度も CD-ROM にアクセスすると応答が遅くなる。また、中間の検索結果集合が大きく膨れ上がる可能性を配慮したソフトウェア作成は、パーソナルコンピュータ上ではなかなか難しい。例えば、中間結果のソートだけをとっても件数が何万件にもなるのであれば、制約されたメモリー空間の中で妥当な応答速度で行なうことは難しい。このような事情のため、良好な応答速度を達成するソフトウェア作成手法の一つとして、主な検索はすべてインデックスの一次検索のみに頼るという方法が考えられる。但し、そのためにはインデックスの作り方に工夫が必要である。

以上のように CD-ROM 検索ソフトウェアの作成に当たっては、PC-9800 のアーキテクチャの制約と CD-ROM 固有のアクセス特性が最大の問題になる。本研究の範囲内では検索ソフトウェアの高機能化等は主要な研究目的ではないので、特にこの点を追求することはしなかった。

PC-9800 上の CD-ROM 検索ソフトウェアの作成については、幸い、大日本印刷が辞書類データベースの作成について多大のノウハウを蓄積しており、CD-ROM のプレスを含めこれを有効に活用する方針で進めた。その主な仕様は、

- (1) すべての分野を横断的に検索できるとともに、一分野に特定した検索も可能とする。
- (2) 検索キーについては、英語、和語漢字形、カタカナ、ローマ字のそれぞれで行なえるようにする。複数単語から成る用語の検索に際しては、英語の場合は空白を入れて普通に入力し、カナの場合はわかち書き等を気にせずに入力できるようにする。

などの点である。

なお、CD-ROM の検索機能の高度化は、エンジニアリングワークステーションなどのアーキテクチャ側からの束縛の少ないコンピュータを対象として検討して行くのが得策であろう。

4. CD-ROM の仕様

CD-ROM の作成に当たっては、先に述べたように原データからインデックス等を切り出すための仕様が主な課題となる。検索を、英語形、日本語形(漢字)、カタカナ、ローマ字の検索キーで行えるようにするには、それぞれのフィールドから然るべきインデックスを切り出し、CD-ROM 上でのファイルとして形成しておかねばならない。しかも少ない回数のアクセスで目的のレコードを見いだすためには、取り出すインデックスの形式について十分な吟味が必要である。原データからインデックスはおおむね次のようにして切り出す。英語、漢字、カタカナ、ローマ字の四つのフィールドのインデックスに分けて考える。それぞれのフィールドについて、例えばデータ値が、

Abc 123 Xyz

のような三つの単語列からなる場合、空白などの文字で切られていたり分かち書きされているところで区切り、単語の組合せを作っていく。この3単語からなるデータ値の場合は、次のようになる。

ABC 一単語

123

XYZ

ABC123 二単語の組合せ

ABCXYZ

XYZ123

ABCXYZ123 三単語の組合せ

これらのパターンは、1ないし3個の単語の組合せをアルファベットの降順に並べ、それを連結したものである。これらをすべてインデックスとして切り出し、一つのインデックスファイルに混在させる。なお、英語、ローマ字はすべて大文字に変換する。カタカナでも小文字を大文字に直し、特に長音記号を取り除く。なお、原データ内の特殊記号については、それぞれの記号に意味にしたがって然るべき処理をほどこす。(データ作成での仕様を参照)

このようにすべてのインデックスの組合せを並べるのは、CD-ROM アクセス回数を少なくし検索ソフトウェアの負荷を軽くするためで、その結果として応答速度の向上を図るなど使い勝手をよくするためである。常識的な情報検索システム流の考え方では、入力されたそれぞれの単語について検索結果集合を作り、集合間で和をとることによって処理するのが普通であろうが、CD-ROM のインデックス検索には1回に0.5ないし2秒程度の時間を要するので、これを避けるための方法として、上述のようなキーワード切り出し方法を採用した。

すなわち、複数の単語を検索キーとして入力した場合は、上述のようなキーワード切り出しと同じアルゴリズムで整形連結処理を施し、一回のインデックス検索のみでレコード番号を得ようとするのである。この結果、単語のAND演算しかできなくなるが、用語検索ではANDが主体で、ORは少ないと判断したからである。ORをとりたいときは、それぞれについて検索をすればす

む。応答が早くそれができれば、大部分の要求は満足され则认为られる。なお、上記の方法では連結の結果偶然にもインデックスのパターンが他の用語と符合すると、検索結果にゴミがはいってしまうことになる。これは、検索負荷を軽くしたことの代償と言えよう。

一方、検索ソフトウェアを高度にすれば、検索結果集合の論理演算も可能になるから、一単語だけのインデックスをまとめたファイルも別途 CD-ROM 上に作成することにした。幸い、CD-ROM は、学術用語の原データ自体がそれほど大きくないために各種のインデックスをいれておくのに十分な容量を持っている。

5. 試作システムの概要

試作した Sci-Term/CD-ROM の特徴は次のようである。

- (1) 21別分野の学術用語集に入っている情報のうち基本的には英和部分の情報をすべて盛り込んであり、全体でおよそ22万8千レコード格納している。
- (2) 外国語、日本語表示形、ローマ字のほかに、新たに仮名読みを付してある。英単語、仮名などがすべてインデックスとして用意されている。従って、用語の一部を用いて検索することができる。(例えば、'computer' を含む用語のすべてを検索するようなことが可能) また、検索する分野は、任意の複数分野を指定したり、あるいは全分野にわたることができる。
- (3) 完成したシステムは学術用語の検索に一般的に使用できるものであるが、特に学会等用語集を新たに作成したり改訂したりする場合に、用語編集者が座右において他分野の用語をチェックするために用いることを配慮して開発を行なった。

Sci-Terms/CD-ROM の検索と結果表示例を図1～3に示す。検索画面上には、外国語、日本語(漢字の用語形)、カタカナ、ローマ字に対応する検索キー入力フィールドを持つ。また、分野の選択も可能で、指定しない場合はすべての分野にわたって横断的に検索を行なう。分野をいくつか指定した場合には、当該分野についてのみ検索を行なう。

検索に当たっては、例えば“abc xyz”という検索キーの入力に対して、まず区切り文字を取り除きつめることにより、ABCXYZ という語をつくる。この際にすべて大文字に直していることに注意されたい。カタカナでも、小文字を大文字に直し長音記号を取り除く。以上の操作はインデックス切り出しの際の処理とちょうど同じである。このようにしてつくられた語により、CD-ROM の検索を行ない、結果を表示する。今後は、さらに論理和、論理積を結果集合の処理によって行なうようなソフトウェアを作成し、その性能を評価する予定である。

結果の表示については、なるべく一覧性のよい画面形式であるよう留意し、操作はファンクションキーを主体にし、マニュアルなしに使えるよう画面表示や HELP 機能を強化するよう配慮した。結果の表示については、図のように用語形のための簡易表示とすべてのデータを出す詳細表示の二通りを設け、用語選定作業の様々な局面で使用できるようにしている。

【図1】

学術用語 CD-ROM

[条件入力]

入力エリア

・複数語句の区切りは 論理積＝スペース、論理和＝“+” です。

検索対象

外国語(用語)	日本語(漢字)	かな(用語)	ローマ字(用語)
外国語(単語)		かな(用語)	ローマ字(用語)

・↑↓←→ で移動し、いずれか1つを選択してください。

対象分野

全分野			
光学	物理学	地学	地理学
電気工学	動物学	原子力工学	地震学
化学	海洋学	計測工学	機械工学
計外工学	航空工学	論理学	採金工学
船舶工学	歯学	心理学	植物学
数学	天文学	図書館学	

・↑↓←→ で移動し、リターンキーで選択・解除して下さい。

ID

【図2】

学術用語 CD-ROM

[結果・簡易]

外国語	日本語	分野
A.C. (Ante Christum【L.】)	西暦(紀元) 前【キリストの降誕前】	キリスト教学
A.D. (Anno Domini【L.】)	西暦(紀元) 【キリストの降誕後】	キリスト教学
a posteriori...【L.】	後天的—〈形〉	キリスト教学
a priori...【L.】	先天的—〈形〉	キリスト教学
abba【Aram.】	アバ(父)	キリスト教学
abbacy	〈1〉修道院長の職【自治修道院の】	キリスト教学
abbacy	〈2〉修道院長の職権【自治修道院の】	キリスト教学
abbacy	〈3〉修道院長の任期【自治修道院の】	キリスト教学
abbacy	〈4〉修道院長の管区【自治修道院の】	キリスト教学
abbess	女子修道院長【自治修道院の】	キリスト教学
abbot	修道院長【自治修道院の】	キリスト教学
abbreviator	教皇庁官房文書速記官	キリスト教学
abiding now	とこしえの今→nunc permanens【L.】	キリスト教学
abjuration	異端放棄の誓い	キリスト教学
ablution	洗淨式【ミサ中の】	キリスト教学
abnegation	克己	キリスト教学
Absolute, the	絶対(者)	キリスト教学

【図 3】

学術用語事典 CD-ROM

[結果・詳細]

外国語	日本語	分野, I D, 出典
A.C. (Ante Christum【L.】)	西暦(紀元)前【キリストの降誕前】 セイレキ (キケン) ゼン seireki (kigen) zen	キリスト教学 CHR810000003 CHR-1981-0105
a priori...【L.】	先天的—〈形〉 センテンテキ... sententeki...	キリスト教学 CHR810000004 CHR-1981-0105
abba【Aram.】	アバ(父) アバ(オチ) aba(titi)	キリスト教学 CHR810000005 CHR-1981-0105
abbacy	〈1〉修道院長の職【自治修道院 の】 シュウトウインチョウ ノ ショク syu^do^intyo^ no syoku	キリスト教学 CHR810000006 CHR-1981-0105
abbacy	〈2〉修道院長の職権【自治修道院 の】 シュウトウインチョウ ノ ショクケン syu^do^intyo^ no syokken	キリスト教学 CHR810000007 CHR-1981-0105

6. まとめ

本稿では、学問分野での用語の和訳を確定する、という作業を念頭に置いた用語検索システムの実例を紹介した。CD-ROM という制約のある環境下でのデータベースシステム実現についての工夫も併せて述べている。今後の CD-ROM データベースの発展・普及を期待したい。

第2章 シソーラスの開発と利用

1. はじめに

シソーラス (thesaurus) の語源はギリシア語で、宝庫、倉庫などの意味から知識の宝庫としての辞書や百科事典などを指す。今日では、データベースの作成および利用に際して使用すべきキーワードの範囲、表記、用法等について定めた典拠として広く用いられている。

データベース作成のための索引づけには、従来から統制語方式と自然語方式の2種類があり、最近ではコンピュータによる自然言語処理の進歩によって、対象となる情報から直接、索引語を抽出する自然語方式が増えてきた。

日本経済新聞社では1975年以来、日本経済新聞をはじめとする日経4紙の記事データベースの構築を進めてきた。この記事データベースの索引づけは、統制語を主体に自然語を併用する方式であり、このためのツールとして「日経シソーラス」を作成、数次にわたって改定を行ってきた。

統制語方式におけるシソーラスの必要性はいうまでもないが、自然語方式においても検索のヒット率の向上のためにシソーラスの果たす役割は大きい。また、日本経済新聞社ではこれまで数値情報の検索には系列コードなどを用いてきたが、シソーラスを用いたキーワードによる検索も可能とするよう開発を進めている。

このように、データベースの利用におけるシソーラスの重要性は今後ともさらに高まるものと思われる。

我が国のデータベースの国際化を図るために日本語情報の英文化が急がれているが、同時にデータベースの利用に不可欠のツールであるシソーラスの英文化も重要な課題である。たとえデータベースが日本語のままであっても、シソーラスを英文化することは日本語データベースの利用の道を一步開くことになるだろう。

本章では以上の観点から、シソーラスの一般的な概要を述べるとともに日経シソーラスの作成基準を対比させ、さらに日経シソーラスの概要を述べることにする。

2. シソーラスの概要

データベースの作成・利用におけるシソーラスの一般的な役割は、索引や検索に用いられる用語に関係のある語を示すことによって、インデクサやサーチャーがより適切な用語を見いだすことができるよう導くことである。このため、シソーラスは以下のような構成および表示方法をとることが多い。

2-1 構成要素、語の関係

シソーラスで規定される語の関係には、通常、次のものがある。

(1) 上位語・下位語：階層関係

A、Bの二つの語があって、AがBより広い概念を表す場合、AをBの上位語(B T = Broader Term)、BをAの下位語(N T = Narrower Term)と呼ぶ。階層関係には次のものがある。

(イ) 属関係：クラス・メンバーの関係で上位語はクラス、下位語はメンバーにあたる。

(ロ) 部分全体関係：からだや機械の全体と部分や部品、地理的關係など。

(ハ) 事例関係：一般名称(上位語)と固有の名称(下位語)

日経シソーラスでは検索の便宜上、上記の外にも階層関係を設定している。

(ニ) 原料と粗製品：(例) 金—金地金、鉛—鉛板

(ホ) 行為・現象とその一部あるいはそれに付随するもの：

(例) 外交—外交文書、経営—営業権、食費—エンゲル係数

階層関係においてはある語が二つ以上の上位語を持つケースもよく起こりうる。これを重合階層関係と呼ぶ。

(2) 同義語：等価関係

複数の用語が同一の概念を表す場合。完全に同義ではないが、索引や検索上で同義としても支障がない用語(準同義語あるいは検索同義語と呼ぶ)も含まれる。索引語としては統制語(ディスクリプタ)が優先的に使用され、非統制語(あるいは非ディスクリプタ)は"USE"の指示で、適切な統制語に導くための「導入語」となる。等価関係は表現の多様性に起因するものであるが、同義語が発生しやすいケースとして以下のものがある。

(イ) 日本語と外来語

(ロ) 略称とフルネーム

(ハ) 俗語・俗称と正式名称・標準名あるいは新語と旧語：

(例) 首切り—解雇、青空駐車—違法駐車

日経シソーラスでは、同義語の内のどれを統制語にするか特に基準を定めていない。

(3) 関連語あるいは参照語：連想関係あるいは参照関係

上位・下位や同義の関係でなく、連想または参照関係にある語を関連語(R T = Related Term)という。この関係は相互的であることを原則にする考えもある。

また、姉妹語(直近の上位語が同一)は一般には関連語として扱わない場合が多い。

連想関係には以下のものがある。

(イ) 部分全体関係：階層関係でなく連想関係にとどまる場合。

(ロ) 学問分野と対象：(例) 地震学—地震

(ハ) 行為・操作と行為者・対象・道具・結果：(例) 禁煙—たばこ

(ニ) 職業とそれに従事する人：(例) 会計—会計士

(ホ) 原材料と製品

(ヘ) 反対概念：(例) 売りオペ—買いオペ

日経シソーラスでは関連性が強い語は階層関係に位置づけている。かわりに参照語(S A = See

Also) を設けて、索引者や検索者がより適切な語を発見できる手引きとしている。SAの参照関係は一方向でもよい。

2-2 表示方法

このような語の関係を表すために、シソーラスは50音順表示または体系表示のいずれかまたは両方をとるのが普通である。日経シソーラスにおいては、両方式を併せた冊子体を作成している。

(1) 50音順表示またはアルファベット表示

見出し語を50音順（またはアルファベット順）に配列して、その語のスコープノート（注釈）およびすべての下位語、上位語の階層関係、さらに関連語（あるいは参照語）を表示する。見出し語が非統制語の場合には“USE”の表示によって統制語を示す。

語の中には、関係する語を一つも持たない場合があり、日経シソーラスではこれを孤立語と呼んでいる。

(2) 体系表示

概念によって語を分類して、それぞれの分類中に最上位語から最下位語までの階層関係のツリー構造を表示する。

3. 日経シソーラス

冒頭に述べたように日経では1975年から新聞記事データベースを作成しており、記事に付与されたキーワードの語彙数は50万語に及んでいる。これらキーワードの管理は、キーワード辞書ファイル、シソーラスファイル、対応語ファイルの三つのファイルを管理することによって行われている。

3-1 キーワード辞書ファイル管理

キーワード辞書は日経の記事データベースに付与されたキーワードの集合で、キーワード辞書ファイルに収録されている。収録項目はキーワードのカテゴリー、漢字表記（2バイト文字）、カナ読み（1バイト文字）、登録日ナンバー（キーワード固有の数値コード）等である。管理の主な内容は、キーワードのカテゴリー分け、表記の統一、略語・略称の管理等である。

◇キーワードのカテゴリー

（語数は1991年1月現在）

会社キーワード		164,384語
団体	//	71,340
人名	//	126,913
品目	//	7,869
業界	//	423
項目	//	6,711

地域 // 433

補助 // 122,651

上記の内、会社、団体、人名各キーワードは固有名詞であり、国内・海外を問わず、記事の主題に関係ある会社、団体・機関、人名を採用する。表記法は原則として日本国内の会社、団体、人名は正式名称、海外はカタカナ書きである。

品目、業界、項目、地域の各キーワードは統制語である。

補助キーワードは統制されない自由なキーワードであり、統制語では表現できない特殊な概念や細かな概念を表現するために用いられる。ただし、主となる索引は統制語であり、主題概念にもっとも近い統制語を付与した上で、その補完として補助キーワードを用いることが原則である。

なお、以上の外に記事中に出現する一般名詞も free term として検索に使用できるよう記事データに付与される。

◇キーワード表記（漢字表記）

キーワードの表記法は新聞表記に準ずるのが原則である。ただし、次のような同型異義語が発生する場合がある。

- (イ) 分野が異なる：(例) ドライバー (ねじ回し、運転手、ゴルフのクラブ)
- (ロ) 英語の略語：(例) CD (キャッシュディスク、コンパクトディスク)
- (ハ) 固有名詞と普通名詞：(例) ライオン (会社、動物)
- (ニ) 固有名詞同士

上記の内、(イ)、(ロ) は言い換え、(ニ) はかっこ書きの限定句 (主に企業名が多いので、業種、所在地など) を付する。

◇略称・略語の管理

会社名などの固有名詞は正式名称あるいはフルネームを索引語として用いるのが原則であり、記事中に現れる略称や略語を正式な表記のキーワードに転換することが必要である。キーワード辞書ファイルにはこれらの略称・略語と正規の索引語の対応関係も収録されている。

3-2 シソーラス管理

日経シソーラスは統制語の管理が中心である。ファイルと冊子体があり、ほぼ3年に1回、定期的にメンテナンスを行っている。

(1) 基本構造

統制語 (品目、業界、項目、地域の各カテゴリー) 同士の階層関係、参照関係を規定。また非統制語の中から、比較的出現頻度の高い語を選んで同義語として統制語と対応づけている。階層関係には以下の制限がある。

(イ) 階層のレベルは6段階。つまり最上位語 (最下位語) は5段階の下位語 (上位語) を持つことができる。

(ロ) 直近下位語の MAX は40語。直近上位語は 5 語。

(ハ) 最上位語から最下位語までの一つのツリーに収録できる語の MAX は200語。ツリーが大きくなりすぎないように本来は上下関係にあるものでも、階層構造からはずす場合がある。

(2) シソーラス・ファイル

統制語の階層関係のみを収録。見出し語をレコード・キーとして、見出し語を中心にした下位語、上位語のツリーを配列として持つ。シソーラス・ファイルは、冊子体の「日経シソーラス」を製作するほかに、サービス用のデータベースを作成する際に、各記事データに付与されたキーワードの上位語を自動的に付与するために用いられている。

(3) 冊子体のシソーラス

シソーラス・ファイルの階層関係に、別途作成した参照語データ、同義語データを加えて作成する。50音順表示および体系表示としてのキーワード分類表（分野別シソーラス）がある。

(イ) 50音順表示

1985年版までは、各見出し語に、下位語・上位語のツリーや参照語とスコープノートを表示していたが、ページ数の増加によってハンディな取り扱いが困難になってきたため、1988年版では分類表を主体にしたものに改めた。50音順表示は索引であり、当該キーワードを含むツリーが掲載されているページを表示している。

(ロ) 分類表（分野別シソーラス）

以下に示す22の大分類と151の小分類に分けて、上位語から下位語へのツリーを表示している。

◇大分類

- | | |
|-------------|-----------------|
| 1 共通 | 12 情報・通信 |
| 2 経済・産業 | 13 建設 |
| 3 経営・企業 | 14 流通・サービス・家庭用品 |
| 4 農林水産 | 15 環境・公害 |
| 5 食品 | 16 科学技術・文化 |
| 6 繊維・木材・紙パ | 17 自然界 |
| 7 資源・エネルギー | 18 国際 |
| 8 金属・土石 | 19 政治 |
| 9 化学 | 20 地方 |
| 10 機械・器具・設備 | 21 社会・家庭 |
| 11 電子電機 | 22 地域 |

3-3 対応語ファイル管理

非統制語（補助キーワード）と統制語の対応関係をファイル（通称「親子ファイル」）に収録し

ている。日経の記事データベースは統制語が主体になっている。記事の主題を表現する specific なキーワードが補助キーワードしかない場合でも、最も近い統制語も併せて付与するのが原則。親子ファイルはこの作業をある程度自動化するために必要なファイルである。

非統制語と統制語の対応関係としては、同義語の外に次のケースがある。(例はいずれも左が非統制語)

(イ) 非統制語が統制語の下位語であるもの：(例) LNG火力-火力発電

(ロ) 検索上密接な関係にあるもの：(例) 放映権-テレビ

(ハ) 非統制語の概念を複数の統制語で表現できるもの

(例) AF一眼レフカメラ-一眼レフカメラ、自動焦点カメラ

3-4 今後の課題

(1) 自然語との対応

記事データベースには索引語として自然語を用いるものが増えてきている。自然語は同義語や類義語などの処理を適切に行わないとヒット率が著しく低下する。また、統制語を用いる場合にも、50音表示や分類表示から適切な用語を見付けだすのではなく、「思い付いた言葉」から容易に統制語に導かれる工夫が必要である。このように、今後のシソーラスの整備は、膨大な数の自然語同士の関係づけや統制語との対応関係を充実させていくことが課題となっている。

(2) 機械による検索支援

日経シソーラスに限らず、これまでのシソーラスは冊子体が主流だったが、よりユーザーフレンドリーな検索のためにはシソーラスをコンピュータに組み込んでディスプレイ上に表示することが必要である。特に、(1) に述べた自然語の同義・連想関係や統制語との対応関係を充実させることによって、利用者が「思いついた言葉」を入力するだけで、同義語や関連語あるいは対応する統制語を表示したり、自動的に検索を実行したりするなどの機能が不可欠である。

(3) 英文シソーラスの作成

日本語のデータベースを英文化するには、完全な機械翻訳が実現するまでは、多くの人手と経費がかかるが、データベースの国際化の第一歩としてまずシソーラスやキーワードの英語化から着手することも必要であり、欧米の利用者からの要望もある。

分野別シソーラスの例（１）

情報・通信一般

情報（品） ・イベント情報 ・医療情報 ・海外情報 ・・カントリーリスク情報 ・価格情報 ・画像情報 ・・図形情報 ・・地図情報 ・・立体映像 ・企業情報 ・・経営者情報 ・技術情報 ・・特許情報 ・気象情報 ・求人情報 ・・アルバイト情報 ・行政情報 ・軍事情報 ・経営情報 ・・人材情報 ・・人事情報 ・・販売情報 ・・・顧客情報 ・経済情報 ・・金融情報 ・結婚情報 ・交通情報 ・・道路情報 ・個人情報 ・・個人信用情報 ・産業情報 ・商品情報 ・信用情報 ・・個人信用情報 ・生活情報 ・地域情報 ・・タウン情報 ・地価情報	・中古車情報 ・統計情報 ・投資情報 ・・株式情報 ・・債権情報 ・ファッション情報 ・不動産情報 ・・住宅情報 ・文字情報 ・・文献情報 ・レジャー情報 ・・観光情報 ・・スポーツ情報 ・・旅行情報 情報化社会（項） ・高度情報社会 通信（品） ・移動体通信システム ・衛星通信 ・音声メール ・画像通信 ・通信回線 ・・公衆回線 ・・専用回線 ・・デジタル回線 ・・・高速デジタル回線 ・・電話回線 ・データ通信 ・・異機種間通信 ・・国際データ通信 ・・データ転送 ・・データ伝送 ・・DDX ・・・回線交換 ・・・パケット交換網 ・・デジタル通信 ・・・高速デジタル伝送 ・・・デジタル通信網 ・・・・ISDN ・・パソコン通信	・・光通信 ・・ファクシミリ通信網 ・テレビ会議 ・テレポート ・電子掲示板 ・電子メール ・電話会議 ・PCM ・プロトコル ・・MAP ・・OSI 通信料金（項） ・テレックス料金 ・電報料金 ・電話料金 ・郵便料金 テレックス（品） 電気通信事業（品） 電報（品） ・国際電報 ・祝電 電話（品） ・クレジット通話 ・航空機電話 ・国際電話 ・コレクトコール ・自動車電話 ・船舶電話 ・ダイヤルイン ・長距離電話 ・テレビ電話 ・転送電話 ・電話予約 ・フリーダイヤル ・ホームテレホン ・ポケットベル ・列車電話 電話番号（品） 郵便（品） ・外国郵便	・・航空便 ・・・クーリエサービス ・切手 ・・記念切手 ・小包 ・私書箱 ・速達 ・はがき ・・絵はがき ・・広告はがき ・・年賀はがき ・ファックス郵便 ・郵便番号 ・郵便振替
---	---	--	---

分野別シソーラスの例（２）

情報産業、ニューメディア（１）

印刷（品）	・住宅情報誌	・電子電話帳	・住宅地図
・オフセット印刷	・女性誌	・同人誌	・世界地図
・活版印刷	・生活情報誌	・ニューズレター	・立体地図
・カラー印刷	・タウン誌	・PR誌	ニューメディア（品）
・曲面印刷	・男性誌	・復刻版	・VRS
・グラビア印刷	・パズル誌	・古本	・AM多重放送
・軽印刷	・ビジネス誌	・ムック	・衛星放送
・平版印刷	・ファッション雑誌	情報サービス（品）	・FM多重放送
・立体印刷	・ミニコミ	・結婚情報サービス	・CATV
広告（品）	・自費出版	・コンサルティング	・CCTV 放送
・インフォマーシャル	・写真集	・情報処理サービス	・Hi-OVIS
・屋外広告	・書籍	・計算センター	・静止画放送
・ネオンサイン	・育児書	・ソフトウェアサービス	・双方向テレビ
・折り込み広告	・絵本	・ソフトウェアハウス	・Hi-OVIS
・カタログ	・教科書	・情報提供サービス	・テレテキスト
・カタログ雑誌	・再版本	・IP	・文字多重放送
・求人広告	・ザラ紙本	・気象情報サービス	・テレポスト
・広告代理業	・参考書	・情報検索	・ハイビジョン
・CF	・時刻表	・情報検索システム	・PCM 放送
・CM	・辞書	・代行検索サービス	・ビデオテックス
・CMソング	・国語辞典	・データバンク	・キャブテンシステム
・車体広告	・電子辞書	・テレホンサービス	・プライベートキャブテン
・新聞広告	・児童書	・電光ニュース	・テリドン
・タイアップ広告	・社史	・シンクタンク	・テレテル
・チラシ	・紳士録	・調査業	・NAPLPS
・DM	・図鑑	・興信所	・ビルトシルムテキスト
・パンフレット	・全集	情報産業（品）	・プレステル
・POP 広告	・専門書	新聞（品）	・ミニテル
・比較広告	・電話帳	・学校新聞	・ファクシミリ新聞
出版（品）	・電子電話帳	・機関紙	・ファクシミリ放送
・機関誌	・年鑑	・業界紙	・ペイテレビ
・コミック誌	・ビジネス書	・縮刷版	放送（品）
・雑誌	・美術書	・スポーツ紙	・緊急警報放送
・科学雑誌	・百科事典	・ファクシミリ新聞	・社内放送
・カタログ雑誌	・文庫本	・夕刊紙	・深夜番組
・クイズ誌	・洋書	製本（品）	・多重放送
・月刊誌	・旅行書	地図（品）	・AM多重放送
・写真雑誌	・電子出版	・案内地図	・FM多重放送
・週刊誌	・電子辞書		
・就職情報誌			

情報産業，ニューメディア（2）

・音声多重放送 ・文字多重放送 ・テレビ放送 ・衛星放送 ・CATV ・CCTV 放送 ・Hi-OVIS ・静止画放送 ・双方向テレビ ・Hi-OVIS ・テレビ番組 ・テレビニュース ・ハイビジョン ・ファクシミリ放送 ・ペイテレビ ・UHF テレビ ・PCM 放送 ・報道番組 ・有線放送 ・CATV ・CCTV 放送 ・Hi-OVIS ・ラジオ放送 ・AM放送 ・AM多重放送 ・FM放送 ・ミニFM放送 ・短波放送 - マスコミ（品）			
---	--	--	--

50音順索引の例

画像応答システム		価値観	245	カップめん	72	カナリア	249
→VRS	146	価値観F	247	カップリング剤	102	かに	70
画像情報	138	家畜	189	合併→企業合併	45	カヌー	235
画像処理	143	家畜飼育	63	かつら	167	カネミ油症	169,233
画像処理システム	143	家畜廃棄物	169	カテーテル	230	カバン	167
画像処理装置		課長	226	家庭	245	かび	190
→イメージプロセッ		かつお	70	家庭医	7,230	かび取り剤	102
サー	140	かつお節	70	家庭医薬品	108	※ 防かび剤	102
画像通信	138	学会	5	家庭F	247	カフェテリア	164
画像データベース	143	楽器	238	家庭教育	228,245	カフェバー	164
画像入力装置	141	楽器店	159	家庭教育F	229,247	株価	30,34
火葬場	153	学校	228	家庭教師	7,228	株買い取り	45
過疎F	244	※ 校舎	153	家庭菜園	63	※ 株売却	51
家族	245	学校移転	16	家庭生活	245	株買い取りF	47
家族F	247	学校移転F	17	家庭生活F	247	株価F	32,35
家族計画	10,243,245	学校給食	164	家庭用サウナ	126	株価指数	30,34
※ 避妊	230	学校教育	228	家庭用精米機	63,119	株価指数F	32,35
※ 避妊薬	108	学校教育F	229	カテコール	101	株価指数先物	40
※ 避妊用品	230	学校教育法	208	家電	134	株価指数先物F	41
家族計画F	11,244,247	学校教育法F	209	家電業界	43	株価見通し	1
家族旅行	238	学校新聞	146	家電業界F	44	株価見通しF	2
可そ剤	102	学校法人	5	家電業界L	43	歌舞伎	173,238
過疎白書	1	学校誘致	10,224	家電量販店	159	株先50	40
ガソノール	81,88	学校誘致F	11,221,225	果糖	71	株式	34
ガソリン	81,86	滑剤	102	華道	173,238	株式F	35
ガソリンエンジン	115	合算課税	24	過当競争	20	株式会社	45
ガソリンスタンド	160	合算課税F	25	※ 販売競争	55	株式買い取り権付き社	
ガソリン添加剤	102	活性アルミナ	100	過当競争F	21	債→フランク債	
カタール	249,250	活性化	16	稼働率	1,20,53		34,51
カタカナ	174	活性化F	17	稼働率F	2,21,54	株式公開	34
家宅搜索	240	活性炭	97,100	カドミウム	94,185	株式公開F	35
家宅搜索F	242	滑走路	160	カトリック	173	株式公開買い付け	
形削り盤	117	カッターナイフ	96,167	カドリニウム	185	→TOB	45
形鋼	93	褐炭	86	カドリニウム・ガリウ		株式先物市場	30,34,40
カタログ	146	褐炭液化	86,88	ム・ガーネット		株式先物市場F	
カタログ雑誌	146	活断層	187	→GGG	131		32,35,41
カタログ販売	55	ガット	191	金網	96	株式市場	30,34
カタログ販売F	57	活版印刷	146	金型	113	※ インサイダー	
型枠	152	割賦販売	55	神奈川	248	取引	240
可鍛鋳鉄	93	※ クレジット	31	金具	96	※ 出来高	1,20
カチオン染料	109	割賦販売F	57	カナダ	251	株式市場F	32,35
カチオン塗装	178	割賦販売店	159	カナダドル	26,196	株式情報	138

第4部 機械翻訳の利用の現状

第1章 G-SEARCHの機械翻訳による海外提供

第2章 日本貿易振興会（ジェトロ）における
機械翻訳利用の現状

第3章 EC委員会における機械翻訳システムの利用

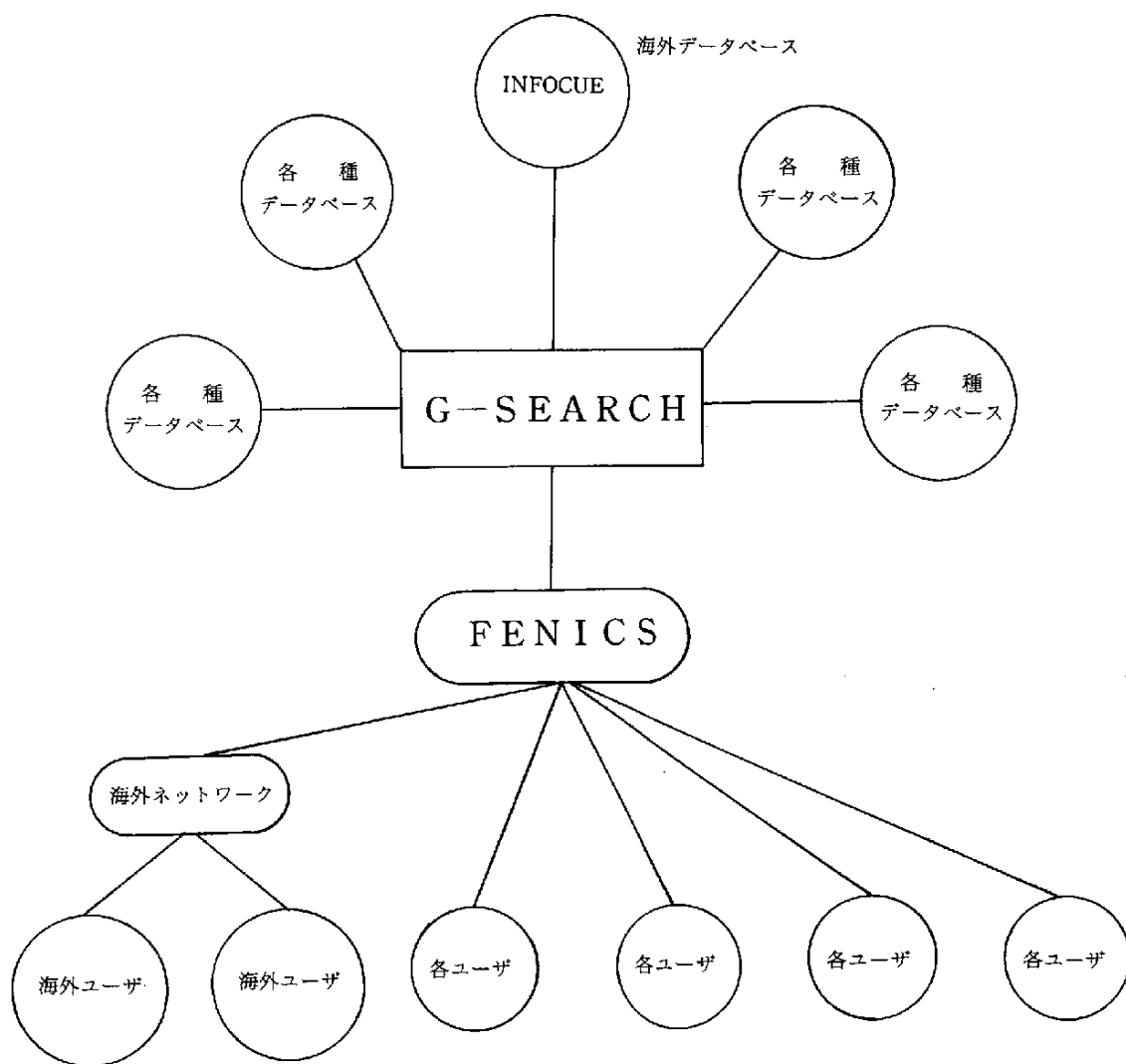
第1章 G-SEARCH の機械翻訳による海外提供

1. はじめに

首記について述べる前に、私共のオンラインデータベースの現状について若干触れてみたい。「より多くの御客様に利用を！」をコンセプトに富士通、富士通 FIP と共に日本初のゲートウェイによる統合データベース「HINET G-サーチ」として国内外にサービスすると共にシステムの改善等いくつかの課題に取り組んできた。

このような状況の中で1991年1月、平和情報センターの一部門であった「HINET 部門」を切り離し、富士通と合併で佛ジー・サーチを設立し、様々な課題に対して各社の技術力、ノウハウを結集して大いなる発展を目指している。

図1 G-SEARCH ネットワーク



2. G-SEARCH の特長

G-SEARCH は企業情報、ビジネス情報、専門情報などの約40の国内データベース・サービス、INFOCUE を通じ約800の海外データベース・サービスが利用可能な、ゲートウェイを利用した相互検索システムである。

2-1 主な特長

- 1) ひとつの ID 番号で G-SEARCH 加入の各センターへアクセスできる。
- 2) ユーザは1度の契約で G-SEARCH の提供する全データベースの検索が可能である。
- 3) ゲートウェイとの特殊プロトコルによりデータベース間的高速切り替えを実現している。

今後も各種の改善を進め、①オンライン問い合わせ対応、②各センター間で異なるシステムの利用方法の統一化などを実施する予定である。

3. データベースの海外提供と問題点

「HINET」時代に開始されたデータベースの海外提供は、その基本的な考え方を継承しつつ、その経験と反省を踏まえ、よりダイナミックに、よりシステムティックに展開される。

但し、現在もマーケットサイズ、ニーズについては把握できていない。阻害している問題、つまり

- ①ターミナルの問題
- ②言語の問題
- ③ネットワークの問題
- ④販売チャネルの問題
- ⑤日本情報の総合的な PR の問題等

も解決には至っていないが、この度富士通開発の機械翻訳システム「ATLAS II」を中心とした外国人用オンラインデータベースシステム「SCANFILE」と G-サーチをリンクさせ、そのプロトタイプを1991年5月フランスで開催される「第3回日本情報に関する国際会議」に出展、説明の運びとなった。

4. SCANFILE 開発の背景

上述の問題点①～②、特に②項に対し、機械翻訳システムを利用することにより問題解決に寄与するところからはじまった。今もデータベースにおいては一部英文化された情報が海外に提供されているが、量・質とも充分とは言えない。しかし人手による翻訳作業はコストが掛かるうえ、時間も要し、データベースの海外提供を阻害している。

他方、機械翻訳システムの研究開発は、多くの大学、研究所ならびにコンピュータメーカーで行われているが完全自動翻訳までには至っていない。

現状では科学技術の関連やマニュアルなどのように対象とする文章の分野を絞って、組織的な翻訳業務の一部で取り入れられはじめています。

翻訳精度については、ここ数年で解析可能な表現がかなり増えてきており、いかに実務の中にもうまく取り入れるか、またいかに使いこなすかがポイントになってきている。従って、データベース検索システムにいかに関リンクさせるかという視点で開発がはじまったのである。

4-1 機械翻訳システム「ATLAS」の概要

富士通では1983年から機械翻訳システムの開発に取り組み、1985年から販売を開始した。その翻訳方式は中間言語方式を採用し、概念構造という表現形式で入力文を理解するのである。

また、ATLASの辞書は各言語に1つあり、解析、生成両方の情報を含み、他言語の辞書から独立している。

各言語に共通の概念記号により、辞書間の対応をとることで多言語翻訳が容易に行えるのである。

4-2 「SCANFILE」の概要

富士通、富士通アメリカ、富士通 FIP の3社は共同でデータベースサービスと機械翻訳システムを統合した機械翻訳利用のオンラインデータベースシステムの実現を目指してきた。

1988年11月の米国の COMDEX では日経テレコムを対象としたデータベース翻訳システムを参考出展し、さらに1989年8月の米国 IFIP (世界コンピュータ会議) でデモンストレーションを実施した。この出展ではコンピュータ関連の記事約1000件を対象に翻訳デモを行い、非常に高い関心を集めた。

このシステムを通じて、米国での関心の強さと潜在需要の大きいことを確信した。以上の経験を基に、本格的な商用サービスに向け検討を重ね、「SCANFILE」の試用システムを作成し、1991年3月より米国でフィージビリティ・スタディを開始した。

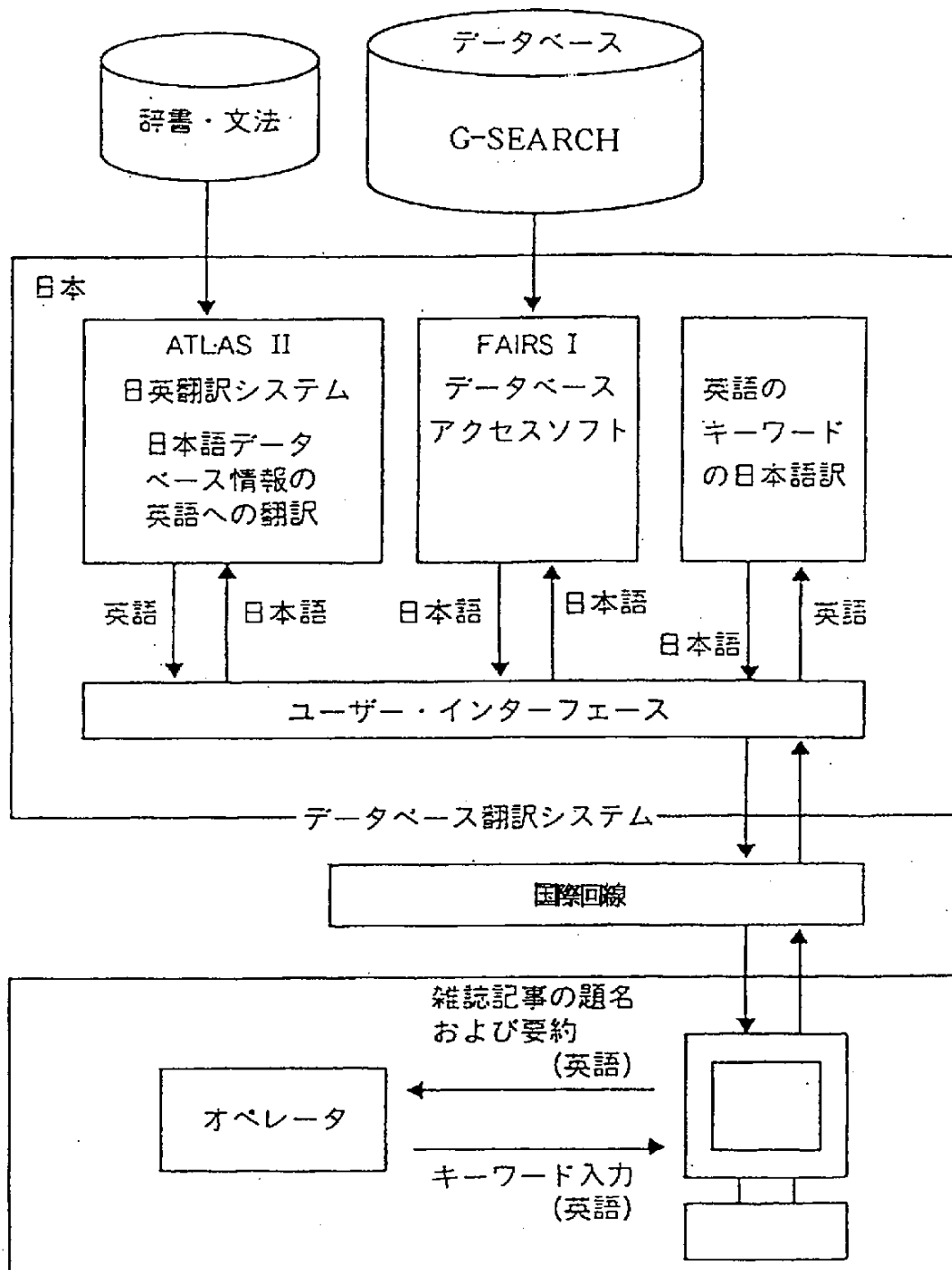
SCANFILEの特長は、日本語のデータベースをあたかも英語のデータベースのように利用できる点にある。

その概要は以下の通りである。

- 1) 英語のキーワードを受け付けて日本語データベースを検索し、検索結果を英文で返却する。
- 2) 機械翻訳を使用して日本語の情報をリアルタイムに翻訳する。
- 3) 正確な訳文が必要な場合は翻訳者による後編集サービスをする。

図2に試用システムの構成を示す。まず、キーワードを英語で入力すると、それを日本語に変換し、日本語データベースをキーワード検索して検索結果(日本文)をリアルタイムに機械翻訳して英文にし、それを画面に表示する。翻訳、データベースなどのすべて

図2 システム構成



の機能は日本側のホストコンピュータに持たせた。その為、端末側の機能は英数字の入出力のみとなり、廉価な端末でも使用可能である。フィジビリティ・スタディでは、各評価ユーザは、米国で広く普及しているパソコン端末を使用している。

また、機械翻訳によるリアルタイム翻訳のシステムとしては、

- 1) 英語のキーワードを日本語に変換する
- 2) 日本語キーワードを検索する
- 3) 検索結果を日英翻訳する、の3つの部分から成る。

1) の課題は単なる英日翻訳では不十分なことである。つまり対象とするデータベースのキーワードそのものにマッピングする必要がある。しかもキーワードはデータベースに特有に付与されている。データベース毎にマッピングする必要がある。その為に、前もって対象とするデータベースからキーワード（日本語）をすべて集め、ATLASを用いて日英翻訳し日本語単語と英単語の対応テーブルを作成しておき、それを用いて英日翻訳をする。

この方法であると各データベース毎に英日変換テーブルを作成することになり、また新たなキーワード追加に伴い頻繁に更新する必要があるが、検索精度を高めるには止むを得ないと考える。

更に、英文によるキーワード検索の機能強化として同義語や類義語への自動展開機能などを付加していくことが今後の課題である。

2) のキーワード検索処理は、現行のデータベース検索と同一である。

3) の日英翻訳では、ATLASを用いて実現している。

いずれにしてもこの部分の翻訳精度がこのシステムの鍵であり、機械翻訳の側面からすると、データベースの文章は、標題文で多く見られる省略表現や、抄録、全文でみられる一文がかなり長い表現など、機械翻訳に困難な表現が数多い。

標題に多く見られる助詞の省略や辻語の省略のある表現については、前もって典型的な標題文の表現の特徴を分析し、文解析文法に特有表現用の文法ルールを追加して翻訳率を上げている。

また長文については、自動的に短く分割することにより、翻訳率を上げると共に、読み安さを向上している。

翻訳精度の向上は今後も継続して努力していくテーマであるが、特に新語の対処が難しい。会社名、人名、商品名などが新語として限りなく発生し、これらを継続的に辞書に登録することは必要であるが、別のアプローチとして、これらの固有名詞を自動的に認定して、そのままローマ字変換するなどの代替案も研究中である。

最後に後編集サービスについて触れておきたい。このシステムでは、検索しながらリアルタイムに翻訳するだけでなく、翻訳者による翻訳の「後編集サービス」を用意している。利用者は、リアルタイム翻訳機能でほしい情報をサーチし、その正確な訳文を必要とする時には、当システムの後編集依頼コマンドを発行することにより、後編集結果を受け取ることが可能である。

このように機械翻訳システムと後編集サービスをリンクさせることによって、より使いやすい、実用的なシステムを可能としている。

4-3 SCANFILE の実用化

本格的な商用のサービスに対しては、まだまだ多くの課題があるがフィージビリティ・スタディを通し、その評価を基に1992年には商用サービス化し、「G-SEARCH」の海外提供に寄与し、弾みがつくものと考えている。

尚、フィージビリティ・スタディの内容については以下を考えている。

1) 日本情報のニーズ調査

2) 機械翻訳の仕組み（システム）に対する評価（品質、価格）

また、スケジュールとしては6ヶ月予定しており、第1フェーズは主に情報に対するニーズ並びに品質・操作性、第2フェーズは市場規模と価格を評価する。

第2章 日本貿易振興会（ジェトロ）における 機械翻訳利用の現状

1. はじめに

日本貿易振興会（以下ジェトロ）において、翻訳（現地語→日本語、日本語→現地語）のニーズは近年増大の一途をたどっている（詳細は後述）。このような中であって、これまでも機械翻訳の活用を検討してきたが、以下に述べるような問題点の解決を待っているのが現状である。

2. 翻訳精度への期待

翻訳後の文書への手直し作業を極力軽減するために、翻訳精度の一層の向上を図る必要がある。

このためには、①適切な辞書の作成、②入力文書記述方法の検討、等の問題を早急に解決する必要がある。

2-1 適切な辞書の作成

これを実現するためには、

- 1) 十分な言葉を含んだ辞書の作成、
- 2) 辞書登録の簡便な方法の開発、
- 3) 利用者固有の辞書作成を可能にする方法の開発等が考えられる。

(1) 十分な言葉を含んだ辞書の作成

このことは単に登録語彙の増加を意味するのみならず、同一文字で意味の異なるものを登録する必要がある。また、熟語、慣用句等を登録して翻訳精度を向上させることも重要である。

また、訳語の出現は Last Use First Out の原則であることが望ましい。

(2) 辞書登録の簡便な方法の開発

メーカー側が作成し提供する辞書のみでは十分であるとは決して言えないのが現実である限り、ユーザーにとってはなるべく簡便な方法によって辞書登録が可能となることが重要である。

(3) 利用者固有の辞書作成を可能にする方法の開発

このことは、上記（2）とも関連するが、翻訳システムをパソコン上に構築し個人で使用するか、ホスト等ワークステーション上に構築し多数で共有するかによるが、現状では後者が一般的である。そこで利用者毎に辞書を作成可能にしておかないと不都合が生じる場合が多い。

2-2 入力文書記述方法の検討

誤訳の少ない文書の記述方法、誤訳が生じ易い文書例等を多数紹介し、ユーザーに予め入力文書作成時に十分注意させるような訓練が必要である。

さらには、誤訳の事例紹介を恥ずかしがらずに、積極的に公開したり、適切な翻訳事例を紹介す

ることも一方法である。

2-3 後処理を軽減させる機能の開発

翻訳精度の向上を図るのみならず、後処理の作業を軽減させる工夫が必要である。例えば、翻訳後の文書をそのままワープロ文書として取り扱える様にする等の工夫が必要である。このことは、後作業を軽減するのみならず、翻訳後文書の浄書機能等多面的な作業の軽減となる。

2-4 翻訳対象文書の機械翻訳機への入力方法の改善

翻訳すべき文書が磁気媒体に記録されているケースが多いとは言えない現状では、文書入力の負担が機械翻訳の利用の妨げとなっている。これを解決するために、OCR 機能の向上等、入力作業の軽減、時間短縮が一方で大きな問題となっており、この面での改善がなされることを期待している。

このような問題点の解決がなされれば、次に述べるようなニーズがジェトロには存在し、非常に活用されるであろうことは論を待たない。

3. 出版物の作成

ジェトロでは、年間相当数の翻訳出版物を作成している。これは、現地語を日本語に翻訳するのみならず、特に近年は、日本語を現地語（主に英語）に翻訳した出版物が急激に増えている。

3-1 現地語から日本語への翻訳

現地事情を紹介した文書、現地の論調等を日本語に翻訳・出版し、国内において現地事情を理解するための一助とする。

3-2 日本語から現地語への翻訳

日本市場の紹介、日本の主張の紹介等を現地語に翻訳・出版し、現地での対日理解を促進する。ここで、「日本語から現地語に翻訳した出版物の作成」における作業の流れは次のとおりである。

- (1) 日本語による原稿作成（手書きが主）
- (2) 現地語への翻訳
- (3) 版下作成・印刷

このように、殆どを人手によっているのが現状である。これまでに、何度か機械翻訳の活用を研究してきたが、実用に至っていない。

今後の方向として、

- (1) 日本語ワープロによる原稿作成、
- (2) 日本語ワープロ文書を入力とした現地語への翻訳（現地語ワープロ）、
- (3) DTP を活用して、現地語ワープロ文書から版下作成・印刷

へと、入力の人手を極力減らすように努力する必要がある。

4. データベースの開発

日本の輸入促進策の一環として、ジェットロでは企業・商品情報、日本市場情報、制度情報を中心に、下記のようなデータベースを開発し、国内50ヶ所の経済国際化センターを通じて提供中である。

近い将来には、国内VAN(複数)を経由して、手持ちのパソコンで利用可能となるような、情報の一般公開も検討中である。

4-1 企業・商品情報

- ★1) 国内の輸入関心企業・商品情報
- ◎2) 海外の対日輸出関心企業・商品情報
- 3) 国内の最新輸入商品情報

4-2 日本市場情報

- ★1) 国内の輸入品売れ筋情報
- ★2) 輸入急増商品の要因分析
- ◎3) 商品別市場紹介
- ★4) 日本で開催される見本市・展示会情報
- ★5) 海外で開催される見本市・展示会情報

4-3 制度、等

- ☆1) 輸入手続き(規則)
- ☆2) 商品別輸入手続き
- 3) 個人輸入関連情報

【注】★：日本語、英語の二重化システム

☆：二重化システムを検討中

◎：英語のみ

無印：日本語のみ

このように、殆どのシステムを日本語・英語と二重化しているのが現状である。(提供情報の端末機でのメニュー画面は図1に示す)

今後の課題として、日本語または英語の入力後、英語または日本語への機械翻訳の活用により、データ更新の迅速化を検討する必要がある。

図1 提供情報の端末機でのメニュー画面

```

*** JETRO-WINDS-BBS ***
-----
<ビジネス情報>
01 ポテンシャル・インポーターズ
02 ポテンシャル・エクスポーターズ
03 最新輸入商品情報
04 アフターサービス情報

11 世界の見本市
12 日本の見本市

<マーケティング情報>
21 日本の輸入統計
22 マーケット・アプローチ(英)
23 マーケット・トレンド情報
24 輸入動向ニュース

<カタログ・ショッピング>
31 海外通信販売カタログ案内

<貿易投資相談>
41 輸入手続き情報
42 商品別輸入便覧

<輸入関連情報DB検索システム>
66 WINDS-DB

<経済貿易情報DB検索システム>
77 WISE

<レファレンス>
98 利用の手引き
99 ジェトロからのお知らせ
-----
番号を指定してください(コマンドEでシステムは終了します)。
?
```

```

*** JETRO-WINDS-BBS ***
-----
<BUSINESS OPPORTUNITY CORNER >
01 POTENTIAL IMPORTERS(OF JAPAN)
02 POTENTIAL EXPORTERS(TO JAPAN)
03 IMPORT GOODS INFORMATION
04 AFTER-SALE SERVICE INFORMATION

11 TRADE FAIRS IN THE WORLD(N)
12 TRADE FAIRS IN JAPAN

<MARKET INFORMATION >
21 JAPAN IMPORT STATISTICS
22 MARKET APPROACH IN JAPAN
23 MARKET TREND INFORMATION
24 IMPORT TREND NEWS

(N): JAPANESE ONLY

<CATALOGUE SHOPPING >
31 MAIL ORDER CATALOGUE GUIDE

<TRADE REFERENCE CORNER >
41 IMPORT PROCEDURE INFORMATION
42 IMPORT GUIDE BY COMMODITY

<DATABASE SYSTEM ON IMPORTS >
66 WINDS-DB

<ECONOMY & TRADE INF. SYS. >
77 WISE

<REFERENCE >
98 USER'S GUIDE
99 INFORMATION FROM JETRO
-----
SELECT A NUMBER (COMMAND E: END OF SYSTEM).
?
```

5. 海外事務所での活用

ジェトロ事務所において、出版物のみならず、講演会・セミナー等を多数開催し、現地での対日理解促進に努めているが、日本から送付された資料・情報を基に、講演会・セミナー用の資料を現地でも作成している。近年では、日本市場に関する情報提供機能の強化のために、海外事務所(38ヶ所、平成3年3月末現在)にマッキントッシュを導入し、現地の対日輸出関心企業に対して、日本の輸入関心企業・商品情報、日本市場の紹介等の情報提供を実施している。このために、種々の情報を日本語と英語で準備し、送付しており、この面でも機械翻訳の実用化への期待は非常に高いものがある。

第3章 EC委員会における機械翻訳システムの利用

1. はじめに

機械翻訳という概念はコンピューターと同じ位古くから存在するにもかかわらず、技術的にも学問的にも関心を持たれることなく見過ごされてきた。機械翻訳が再認識されてきたのはこの2、3年のことであるが、それは数こそ少ないが、実際に成功している実用例によって証明されている結果に負うところが大きい。

このかなり急激な変化に対する説明としてはどんな事があげられるだろうか？ 第一に需要の存在があげられる。翻訳に対する需要は毎年約15%の割合で増え続けており、ヨーロッパ諸国の言語ばかりでなく、日本語、韓国語、アラビア語、中国語といった言語間の翻訳に対する需要も盛んである。

次に要求される翻訳物のタイプに変化がおこってきた。50年代と60年代には、文学作品の翻訳が主流であったが、70年代、80年代に入ると、技術革新が進む中で、技術情報が最優先される様になった。そして最後に国際政治、殊にヨーロッパにおける政治の変化により、新しいマーケットが開かれ、多種言語によるドキュメンテーション及びコミュニケーションに必要な商品とサービスが益々多岐にわたって求められる様になってきた。

現在使用されているシステムのほとんどは60年、70年代にアメリカで生まれたものである。それらは基本的に二つのカテゴリーに分けられる。

すなわち、Logos, Spanam, Systran の様に、より大規模で複雑なシステムは通常、中央集中型メインフレームのコンピューターに設置されており、通信によるアクセスが可能である。一方、Smart, Weidner, Linguistic Products といった小規模でシンプルなシステムは、ユーザーの現場にあるパーソナル・コンピューターにより運用されている。

翻訳の品質は、機械翻訳に内蔵されている言語の組み合わせ、ドキュメントのタイプ、そして言うまでもなく専門用語の網羅性の度合によって左右される。ある翻訳物が、ある言語間の組み合わせにとっては妥当なレベルとなるが、他の言語間の組み合わせにとっては、より不満足な結果となるといった事は往々にしてみられる。

機械翻訳システムのカバーしている言語は現在のところ良好といえる。機能しているシステムのほとんどはフランス語と英語の相互翻訳機能を持っている。又、ドイツ語とスペイン語はソース言語若しくは対象言語としての機能を持ち合わせている。

英語は、ソース言語として最も高度に開発されており、対象言語として、スペイン語、オランダ語、ポルトガル語、デンマーク語、スウェーデン語、日本語、アラビア語といった数多くの言語を持っている。

2. 機械翻訳の利用者

機械翻訳の利用者は現在大きく二つに分類される。それは、情報入手のために外国語の文献を調べる必要のある人々と自国語の資料を提供するために各種の対象言語に翻訳したい人々である。

機械翻訳システムを活用した情報調査の草分けは、アメリカ空軍である。空軍では1970年以来、Systran システムを利用して最初はロシア語から英語に、そして後にフランス語とドイツ語から英語に翻訳してきた。翻訳物は、広範囲の技術分野を網羅しており、空軍はその翻訳レベルに満足しているという。ヨーロッパでは、西独の Karlsruhe にある原子力研究センター (the Nuclear Research Centre) が同じ様な機能を有しており、フランス語の研究論文を英語に翻訳している。どちらの場合においても、機械翻訳されたままの原稿が、主題別分野の専門家によって校正されるようになっている。

第二の利用グループによる機械翻訳の最も一般的な活用例として、メンテナンス・マニュアルの翻訳があげられる。機械翻訳システムのほとんどは、この分野で利用されており、利用者としては Xerox, IBM, Dornier, Siemens の様な大会社が上げられる。そこでは、かなり高度な品質が要求されており、それを達成するために、ソース・ドキュメントの入念な準備、機械翻訳システムの技術用語の信頼性確保及び最後に人による手直しを組み合わせている。利点の主なものは、スピード (製品は外国のマーケットへより早く公開する事が可能)、一貫性とコストが上げられる。

公共部門でも、NATO、国連機関や EC 委員会も、機械翻訳を利用して技術レポート、行政資料、会議議事録の翻訳を行なっている。機械翻訳のみでの品質は、利用者の要求を満たすには至らないが、迅速な後処理 (一時間 4 ページの割合) により、満足な品質にする事が出来る場合もある。

最後に、フランスにおいては、Minitel ネットワークで、機械翻訳がかなり利用される様になってきている。Gachot S. A.により、Systran システムを利用したのオンライン・サービスが提供されている。

3. 機械翻訳研究の現状

ヨーロッパにおける主要な機械翻訳関連のプロジェクトとしては、欧州共同体とそのメンバー国による Eurotra がある。ヨーロッパの言語全てに対するパイロット・システムが1990年末までに完成する予定であったが、一年間延期され、1991年末の完成を目標に現在研究活動が続けられている。

ヨーロッパにおけるその他のプロジェクトとしては、オランダの DLT (Distributed Language Processing) と、当初は Texas から、現在はドイツの Siemens によりサポートされている METAL が上げられる。

4. EC 委員会の Systran 翻訳システム

現在機械翻訳は、インターナル・ネットワークに接続されている Unix を使用できる委員会の誰もが利用可能である。

現在のところ、毎月平均200件にのぼる Systran を使った翻訳依頼があるが、それは翻訳ソフトウェアが搭載された Amdahl 社のメインフレーム・コンピューターとリンクしている高性能のインフラストラクチャーである UNIX ネットワークのたまものである。

これら依頼のほとんどは、いわゆる Systran Raw Service (人的介入なしの機械翻訳) を利用している部局から直接来ており、一方その他の依頼は、さまざまな手直しを行なっている翻訳部門から来ている。実際に、ブリュッセルとルクセンブルグにある 6 つの翻訳部門は、急激に増えつつけている仕事量をさばいていくためには、SYSTRAN をどの様に活用すべきか検討している。

利用者に対する回答時間の平均は通信も含めて約12分となっているが、比較的短いドキュメントの場合は 3 分～4 分以内で済んでいる。

翻訳の質は、言語間の組み合わせ、対象分野、ドキュメントのタイプによって、かなりなばらつきがみられるものの、利用者への調査によれば、このサービスに満足している利用者もぼつぼつ出てきているようである。

ではここで、委員会における Systran の歴史をたどり、その開発と応用面のキーとなっている諸点について論じる事とする。

1976年以来、委員会では Systran を開発してきているが、それはもともとアメリカでいくつかのヨーロッパ言語の翻訳用に開発されたものであった。現在38人の職員と契約スタッフからなるチームが12組の翻訳システムの開発作業を進めている。すなわち、英語から 6 ヶ国語へ (フランス、イタリア、ドイツ、スペイン、ポルトガルの各国語)、フランス語から 4 ヶ国語へ (英国、ドイツ、オランダ、イタリアの各国語)、ドイツ語から 2 ヶ国語へ (英語とフランスの各国語) となっている。

これまでの経験によれば、ラテン系の言語と英語は、ゲルマン系言語に比べて、機械翻訳開発に際して問題となる点が少ない。その結果、英語とフランス語、フランス語とイタリア語といった組み合わせの翻訳の出来具合の方が、ドイツ語からフランス語、フランス語からオランダ語へといったものよりかなり優れたものとなっている。前者の様な組み合わせと同様な翻訳レベルの向上が全ての言語の組み合わせで力強く進んでいる。

5. 言語上の開発

Systran の委員会版は、近年来モジュラー化が大いに進んできている。ソース言語分析は、基礎的辞書に関してさえも、今や翻訳対象から完全に独立している。例えば、英語分析の結果がそれぞれ別の 6 つの対象言語への翻訳に利用されている。また対象のレベルにおいても、同様の総合

プログラムが異なるソース言語に関しても大した変化なく利用されている。このシステムの唯一、正確に二ヶ国語併用可能な部分は、効果的な文脈辞書によるコーディングが、語句若しくは文脈構造の正確な翻訳の鑑定に利用されているトランスファの段階に存在している。しかしながら、この部分においてさえも、一つの決まったグループの一言語（例えば、ラテン系の一言語）に適している文脈上のルールは、同じグループのもう一つの言語への翻訳にも往々にして利用可能であり、それも、ほとんど修正される事なく又修正皆無の場合もあるという事が判明している。

Systran には高度に開発された構文法と語義に関するものを処理する特質が備わっているために、十分に開発された言語の組み合わせに対しては、文法上の homography（一字一音主義のつづり法）と polysemy（多義性）という古典的な問題はほぼ解決されている。この様にして、品質改良の基礎が、特定主題別分野やドキュメントの種類が必要とする用語を組み入れる事によって対象言語へのすぐれた変換対象に基づいている。

現在 Systran の英語とフランス語用の辞典には約 8 万の収録語があり、成文テキストに出てくる言葉のほとんどが網羅されている。しかしながら、文脈辞典には現在 2 万から 8 万の収録があり、その数はたえず増加している。現在の見通しとして、主だった言語の組み合わせにとって、最終的には 20 万の収録を持った文脈辞典が必要になると考えられている。

6. 技術的基礎構造

委員会の Systran は MVS OS の Amdahl 社製メインフレーム (IBM 互換性) 上で作動している。一時間に 40 万語の割合で、翻訳作業がなされている。内部ネットワークによるアクセスは、Unix サーバーを経由しており、ブリッセルとルクセンブルグにある委員会のさまざまな部局からの翻訳要請に応じている。利用者にとって使い易いメニュー・システムによっており、現在 400 以上のワークステーションで利用することができる。その他、要請のあり次第接続が可能となっている。

利用者のほとんどは現在 Unix WS か MS-DOS パソコンを有しており、それらは後処理のためばかりでなく、ドキュメントの作成にも使われている。現在 DGX III でテスト中の OCR は機械可読なドキュメントの作成にとって、大変有効なことが証明されている。

理想的には、利用者はワープロと通信との両用のソフトウェアを利用すべきである。もっとも、ほかのソフトウェアからテキスト変換は可能ではあるが。ワープロソフトである MS-WORD 付きの MS-DOS マシンが Systran 環境の中へ組み込まれる様になる日が早い事が望まれる。

7. 後処理

利用者が修正なしの粗翻訳で満足する場合もあるが、多くの場合機械翻訳した事後の手直しが必要となる。殊に利用者がドキュメントを自分自身の情報のために翻訳させた場合にはなおさらそうである。

委員会の翻訳者は、この問題についてはそれぞれの方法で取り組んでいる。ある方法としては

伝統的手法により最終的なものを作成するためのベースとして機械翻訳を利用しているし、他の方法としては、書類上若しくはスクリーン上で手直しをやっている。いづれにしても、利点・難点があり、個々の翻訳者の経験にとりわけ深くかかわっている。

委員会の要求に応じられる Systran の可能性を評価するため、翻訳者のグループが結成され、各種類のドキュメントから得た結果を調査する事になった。そしてその際念頭におく事として、利用者から要求される質、タイム・リミット、翻訳の目的などがある。同時に、粗翻訳のための Systran へのアクセスをより一層一般化するべく努力がなされており、粗翻訳は従来の翻訳サービスがさし迫った期限内に不可能な場合とか、手許のドキュメントが人手によって翻訳する程重要性がないと考えられている場合にとりわけ効力を発揮する。

8. 利用者

委員会が Systran を開発する主要な目的は内部での利用を促進する事にあるが、我々が獲得した利用者としては EC のメンバー各国における公共団体を網羅している。

現在のところ、最も重要な外部利用者として、ブリュッセルの NATO、ドイツのカールスルーエにある原子力研究センター(Nuclear Research Centre)やドイツ鉄道があげられる。ここ数年来こういった利用者及び他の機関の貢献によって数多くの専門分野別の Systran 辞典及び言語ソフトウェアの拡大がなされてきた。この様にして翻訳の質が、全般的に向上してきている。

The Gachot グループは、民間セクター向けの権利を所有しており、共同体の言語全てを対象とした開発の協定を委員会と結んでいる。

その為、パリ近郊の Soisy-sous-Mont morency にある Gachot S. A.は、関心のある関係者全てがミニテル端末か機械翻訳の要求により良く適合するネットワーク経由で Systran にアクセス出来る様にする立場にある。

最後に、委員会は英語、フランス語、ドイツ語、スペイン語への特定翻訳に関する大きなプロジェクトを準備中である。我々としては、新しい技術に関する専門用語統合のためのベースとしてヨーロッパ特許庁と協定を結ぶ意向を持っている。

9. 将来展望

今後12ヶ月にわたって、委員会は第4のソース言語としてスペイン語を英語に翻訳するための開発にとりかかるだろう。又、ギリシャ当局の要請で、ギリシャ語を英語からの対象言語として組み込む事になるかもしれない。

より一般的なレベルにおいては、我々は現在文脈上のコーディングの分析向上及び情報追加の手段として語義コーディングのレベル向上の開発を研究中である。ソース言語としてのドイツ語のために、複合名詞をそれぞれの構成部分に分類するための特別のてだての開発に取り組んでいる。その目的はそれ自身の構文法と語義に関しての文法的な分析がなされるためである。

企業内の技術的なインフラがより一層安定してきたことから、目標を定めた膨大なドキュメン

トのフォーマット化の可能性を検討すべきであろう。今日までは、ブランクとキャリッジリターンだけをベースとした修正可能なプリントアウトのイメージを利用する他はなかった。それは多種に及ぶワープロやパソコンの機器がインターフェイスなしで Systran 内へ接続されていたからである。

そしてなканずく重要な事としては、開発チームはこれからもひき続き、利用者である翻訳者又は Systran を利用している依頼人から受けた問題解決への回答を出来る限り効率的に行っていく事だろう。それはコンピューター化した言語翻訳の理論的というよりも実際上の問題へのシステムによる応答を確実にしていくためである。

代名詞の指示や、なめらかな文体に留意した表現といった数多くの問題の完全な解決は当面みられそうにもない。

しかしながら、より多くの技術用語とより良い文脈上のルールを加える事によって、急速に増えている利用者が満足出来る翻訳レベルに到達出来る事が望まれている。

10. 結 論

Systran を使った翻訳の質が、すでにかなり満足しうるレベルに達しているのは、フランス語から英語、フランス語からイタリア語、英語からフランス語、英語からイタリア語、英語からスペイン語の組み合わせであり、目下開発中の他の言語の組み合わせについても同様なレベルにまで達する事が望まれている。

委員会が最優先としているのは内部のニーズに対する開発であるが、それはメンバー国家の政府団体のみならず他の共同体機関と利用協定を結ぶ事も出来るのである。

将来における機械翻訳の質的な向上は、プロジェクトに任命されている企業内の翻訳者グループとの共同作業に当たっている各分野別の利用者の要求に徹底的に基づいたものとなるだろう。このような質的向上はより良質の辞典と Systran ソフトウェアに含まれているより充実した意見論的特徴によって一層促進されるべきである。

国際問題委員会報告書

発行日 平成3年3月

発行 財団法人 データベース振興センター

〒105

東京都港区浜松町2丁目4番1号

世界貿易センタービル7階

TEL 03-3459-8581

印刷所 システムワールド株式会社

〒105

東京都中央区日本橋堀留町1丁目11番10号

TEL 03-3639-2560

(禁無断転載)

