自然言語処理　46－6
（1984. 12. 7）

Muプロジェクトにおける総合システムの基本設計

坂本義行
（電子技術総合研究所）

有賀妙子
（ファコム・ハイタック株式会社）

1．はしがき

機械翻訳システムの性能を決定するのは、言うまでもなく翻訳ソフトウエアと辞書である。一方、実用化システムとして利用される際の使いがっての良さを左右するのが、翻訳ソフトウエアと利用者とのマン・マシンインターフェースに当たる部分である。この部分は基本的に翻訳の実行、テキストおよび辞書の編集・管理といった機能を持ち、機械翻訳システムの顔といえる。

科学技術庁機械翻訳プロジェクト(Muプロジェクト）では、言語処理システム・辞書システムとともに、実用化の運用を目指して、顔に当たる総合システムの開発を行っている。この3システムを合わせた機械翻訳システムの概念図を図1に示す。

総合システムを含む日英翻訳システムは、60年3月に工技院情報計算センタ(RIPS)にて、稼動の予定である。ここでは、RIPSでの稼動を目的に開発している総合システムの基本設計について述べる。

2　基本構想

機械翻訳システムの利用方法の典型的なものとして次の二通りが考えられる。

(1)　多量文章の一括翻訳

バッチ処理的に多量の文章を翻訳し、その後集中的に後編集を行う。

(2)　論文の作成・翻訳

日本語エディタ等で作成した文章を翻訳し、原文・訳文ともに画面上で編集する。

この二通りの利用形態を実現するために総合システムは、次のような機能を持つ必要がある。

(1)　原文・訳文の同時編集

(2)　対話的な翻訳依頼

(3)　バッチへの一括翻訳依頼

(4)　汎用の日本語エディタ・英文エディタとのインターフェース

また、さらに使いがっての向上を考慮し、

(5)　文献情報検索システムとのインターフェース

(6)　辞書の編集・管理

(7)　文章（対象原文）に合わせた辞書の選択

(8)　翻訳テキストの管理

といった機能を合わせて備える。

なお、本システムの利用者としては、言語処理システムについて全く知識のない者を想定している。この点を考慮し、利用者が翻訳作業を効率良く行えるよう操作性の向上にポイントをおいた。

3　システムの構成

総合システムは、大きく四つのサブシステムに分けられる。

(1)　翻訳エディタ

原文・訳文双方の対応を取りながらの編集

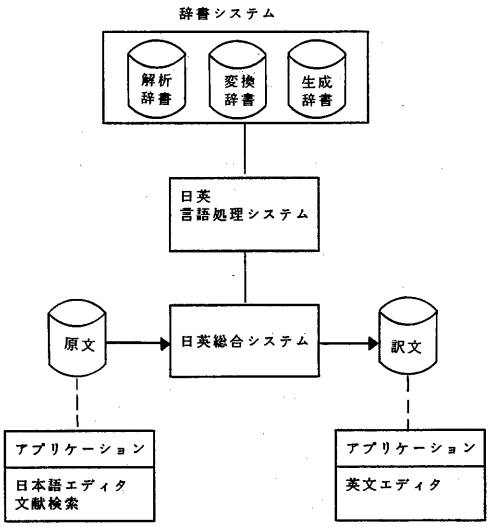

図1．日英翻訳システムの概念図

本研究は国の科学技術振興調整費による「日英科学技術文献の速報システムに関する研究」の一部として行ったものである。その遂行のために総合システム作業分科会を組織し、その審議、指導のもとに研究を進めている。

(2) 辞書エディタ

辞書の内容を対話的に編集

(3) 入出力テキスト変換

外部日本語エディタ・英文エディタおよび文献検索システムとのインターフェースに当たる。それらに固有なファイル形式の入出力を可能とする。

入力変換は、原文ファイルの文章を一文ずつ切り出し、翻訳ファイルへ格納する。また、出力変換は、翻訳ファイルの訳文を指定された形式の訳文ファイルへ書き込む。（翻訳ファイルについては4.2）

(4) 制御・管理

言語処理システムと総合システム間の制御・対話制御・一括翻訳依頼・ファイル管理等を行う。

各部の関連図を図2に示す。総合システムが実現している種々の機能は、制御・管理サブシステムが表示する初期メニュー（図3）より、呼び出す。

4 翻訳エディタ

翻訳エディタには、メニューとエディタの2つの画面がある。（図4・図5）メニューで指定した翻訳ファイルの内容が編集の対象となる。

4．1 辞書の指定

使用する辞書をメニューで指定すると、言語処理システムは、翻訳の際その辞書を参照する。このよ

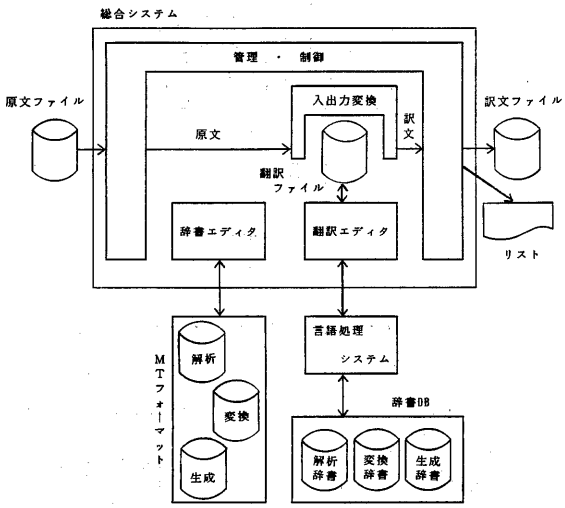

図2．総合システム各部の関連図

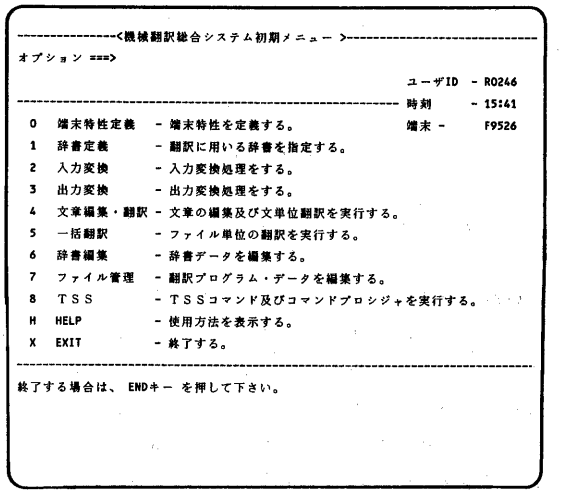

図3．初期メニュー

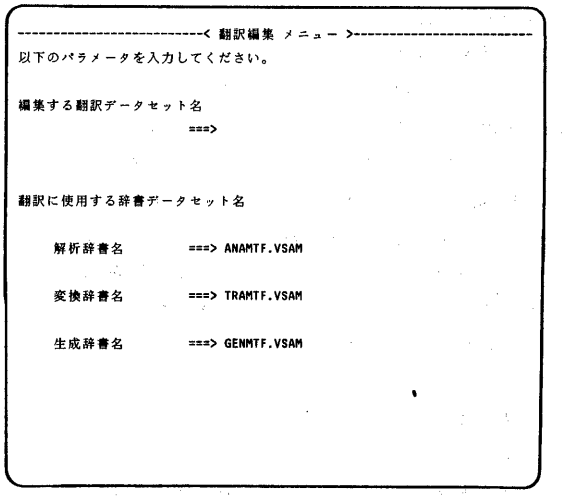

図4．翻訳エディタメニュー

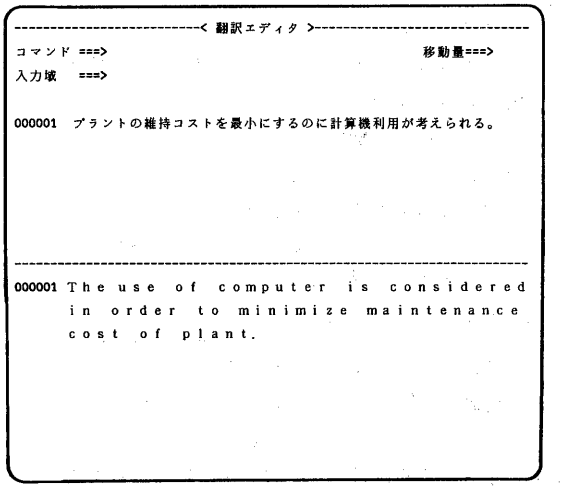

図5．翻訳エディタ

うに翻訳対象の文章の種類によって、最適な辞書を選べる。

4．2　翻訳ファイル

翻訳ファイルには、原文と訳文が文番号ごとにペアになって格納されている。ここで文番号とは、入力変換が文区切りの際に原文ファイルの先頭文から順につけた番号である。

原文と訳文を一文ごとにペアとして、一つのファイルに格納しておくことにより、編集の際の対応がとりやすく、また後編集作業を中断するような時にも、利用者が原文データセットと訳文を格納したデータセットの対応に気を使う必要がない。

4．3　文章の表示

(1)　表示方法

文章を文単位に文番号付きの形式で表示する。原文は上部に、訳文は画面下部に表示され、文番号により、原文と訳文の間の対応が取られる。ここで、文単位とは入力変換によって切り出された一文のことをいう。

翻訳エディタに切り出した文を一文ずつ表示するのは、利用者が見やすくなるということの他に大きな理由がある。それは、利用者が自動文切り出しの結果を確認できるという点である。一文の認識は機械翻訳において大変重要なポイントとなる。画一的な自動文切り出しを避けたい場合も多い。そこで、文切り出しの済んだ原文を表示し、エディタ上でチェック・修正できるようにした。これにより総合システムは、利用者の判断を加味した正しい一文を言語処理システムへ渡す。

文を分割・結合・削除・挿入するコマンドが用意されており、これらを実行した場合には文番号がコマンドの処理に応じて変更する。原文表示域・訳文表示域いずれか一方で文番号に変更があると、他の表示域の文番号も同様に変更される。

翻訳ファイルに原文がなく、端末より原文を入力する場合には，文番号は１から順に振られる。

(2)　禁則処理

表示の際、行頭または行末にきてはいけない文字（記号）を行頭・行末に位置しないように割り付ける。禁則処理の対象となるのは三種あり、表１に該当するものを示す。

文字	対象文字・記号	備考
行頭禁則文字	、。，．？　）：； ブランク	英文に限る
行末禁則文字	（ 単語の途中	英文に限る
ぶらさげ文字*	、。，．	和文に限る

＊　行頭禁則文字の中で行末にぶらさげ出力できる文字

表１．禁則対象文字

(3)　表示の移動（スクロール）

スクロールキーによって、表示が移動量分だけ前後に移動する。またMAXコマンドにより文章の先頭または終端へ移動する。LOCATEコマンドでは、オペランドで指定した文番号の位置へ表示が移動する。

なお、これらの移動は原文・訳文ともに行われ、画面上には常に対応する原文・訳文が表示される。

4．4　文章の編集

文章表示域の文字列を直接修正する。入力はカナ漢字変換またはローマ字漢字変換により行う。

(1)　置換

置き換えたい文字にカーソルを合わせ文字を入力する。

(2)　文字の挿入

挿入したい位置の直後の文字にカーソルを合わせ、挿入モードキーを押し、文字列を入力する。

(3)　文章の入力

翻訳ファイルに文章が存在する場合は、INSERTコマンドを挿入したい文位置の直前の文の文番号部分に指定する。すると、空白文が作られるので、そこに直接入力する。文章がない場合には空白画面が表示されるので、文単位に自由に文章を作成できる。

(4)　文字の削除

削除したい文字にカーソルを合わせ、削除キーを押す。

(5)　文章の削除

削除したい文の文番号の部分にDELETEコマンドを指定する。

4．5　コマンド

上にすでに述べたコマンドの他に編集に役立つコマンドが用意されている。その一覧表を表２に示す。

コマンド名		機　　能
入力行コマンド	MAX	スクロールの最大を指定
	LOCATE	指定する文番号の文を画面の先頭に表示
	FINDJ	文字列を日本語表示域から検索、表示
	FINDE	文字列を英語表示域から検索、表示
	CHANGEJ	日本語表示域の文字列を他の文字列に変換
	CHANGEE	英語表示域の文字列を他の文字列に変換
	CANCEL	編集結果を無効とする
行コマンド	DELETE	文を削除
	INSERT	文を挿入
	JOINT	文を結合
	REPEAT	文を繰り返す
	SEPARETE	文を分離
	TRANSLATE	文を翻訳

表2．翻訳エディタコマンド一覧

4．6　翻訳

翻訳したい文の文番号に翻訳コマンドを指定すると、その文が言語処理システムに渡る。翻訳が終了すると、訳文表示域の対応する文番号の部分に訳文が表示される。

翻訳対象文は、文番号ごとに一文として言語処理システムに渡される。よって、同一文番号中に二文以上の文を含むと正しい結果が得られない。翻訳エディタには、文切り出しの終了した原文が表示されているはずであるが、文分割コマンドや文結合コマンドの実行で、一文でなくなる可能性もある。この点については、利用者が注意する。

5　辞書エディタ

辞書エディタにはメニューとエディタの二つの画面がある。（図6・図7）メニューで編集に関連するデータセット名を指定する。

5．1　辞書システム

翻訳で用いる辞書の基となるデータはJICSTで作成され、MTフォーマットデータと呼ばれるデータ形式（S式）で提供される。辞書システムでは、これを品詞と見出し語をキーとしたVSAMファイル（以後MTF-VSAMと呼ぶ）に変換する。さらに言語処理システムが翻訳に使用する辞書DBには、MTF-VSAMのLexicalな情報の他に辞書ルールが付け加えられている。辞書ルールは、翻訳の変換・生成過程において必要な語固有の手続き的な情報をMTF-VSAMから自動生成したものである。よって普通の意味で"辞書"に当たるのはMTF-VSAMであるといえよう。

なお、MTF-VSAMには、解析・変換・生成の三種がある。

5．2　辞書エディタの基本構想

(1)　編集対象

辞書エディタはMTF-VSAMを直接の編集の対象とする。これには次の理由がある。

①　利用者が辞書を編集する場合修正・追加するのはLexicalな情報に限られる。

②　辞書ルールにはLexicalな情報に基づいて生成されるため、もし辞書を直接編集対象とする

```
----------< 辞書編集 メニュー >----------
処理を指定して下さい。===>          (1：編集，2：印刷)

1．辞書の編集をする場合
    MTフォーマットの種類 ===>    (1：解析，2：変換，3：生成)
    MTフォーマット名     ===>
    マネジメントファイル名 ===>
    更新データファイル名   ===>
    辞書データセット名     ===>




2．辞書内容を印刷する場合
    印刷するMTフォーマット名
                         ===>
  JCL
  ===>
  ===>
```

図6．辞書エディタメニュー

```
----------< 辞書エディタ>----------
コマンド ===>                              移動量===>
見出し語 ===> わたる
品詞     ===> DOUSHI

((SEQ 314)
 (J_LEX わたる)
 (J_CAT 動詞)
 (USAGE ((J_SURFACE_CASE (J_SURFACE_CASE1 が))
         (CASE-PATTERN V1)
         (J_DEEP_CASE (J_DEEP_CASE1 主体))
         (CONDITION ((E_LEX ｒａｎｇｅ１)))
         (A (E_LEX ｒａｎｇｅ)
            (E_CAT V)
            (E_SURFACE_CASE (E_SURFACE_CASE1 ＳＵＢＪ))
            (CASE-PATTERN V1)
            (E_DEEP_CASE (E_DEEP_CASE1 ＯＢＪ))
            (CORRESPONDENCE (CORRESPONDENCE1 主体))))))
```

図7．辞書エディタ

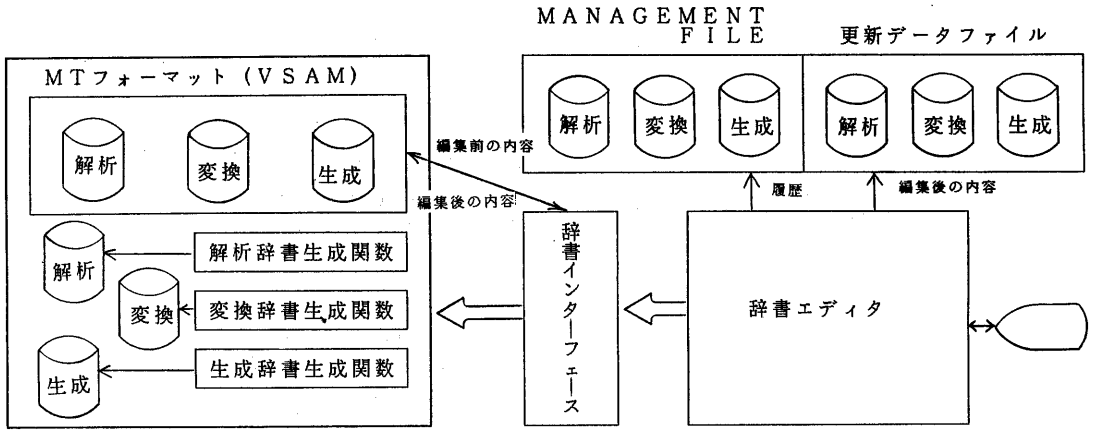

図８．辞書編集におけるデータの流れ

と、エディタで修正・追加したLexicalな情報が辞書ルールに反映しないことになる。

③ MTF-VSAMは品詞と見出し語をキーとして直接アクセスできる。

しかし、MTF-VSAMを編集するだけでは、編集結果は辞書DBに反映しない。そこで辞書エディタを終了する際に、エディタで修正した語の辞書DBを辞書システムの辞書作成関数を用いて作成する。図８に編集作業におけるデータの流れを示す。

(2) MANAGEMENT FILE

MTF-VSAMと同じキーを持つVSAMファイルで、辞書エディタで編集した語についての履歴を格納する。履歴の内容はJICST提供のオリジナルMTフォーマットデータと異なっているか否か、および更新・追加・削除の年月日時分秒である。

MANAGEMENT FILEにはこの他、更新、追加、または削除した語にその旨を記したフラグが記録される。辞書システムの辞書作成関数はフラグを参照し、その語のMTF-VSAMの内容から辞書DBを作りなおす。このフラグは一時的なもので同一辞書エディタセッション中のみ保存される。

(3) 更新データファイル

JICST提供のMTフォーマットデータと同一の形式であり、編集した結果が格納される。

上に述べた点までで、利用者は任意に辞書を変更できるが、それは辞書の台帳ともいえるJICSTデータとは無関係である。各利用者が別個に編集した内容は、今後辞書を拡充していく上で、貴重なデータであり、何らかの方法で収集し、取り込んで行く必要がある。更新データファイルはこの点を考慮して、JICSTへのフィードバック用に用意されているものである。

５．３　辞書内容の表示

辞書エディタ画面の見出し語入力域に参照したい語の見出し語と品詞を入力する。メニュー画面で指定したMTF-VSAMが検索され、ヒットした語の内容がMTフォーマットデータの形で表示される。未登録語の場合は、その旨表示される。

見出し語のみを指定すると、対象としているMTF-VSAMに存在する全品詞で順に検索し、登録されている品詞のリストを表示する。いずれかの品詞を選ぶとその内容が表示される。

５．４　辞書内容の編集

辞書内容表示域内の文字列を直接修正する。辞書の内容はEBCDIC・日本語(JEF)コード混在のため、入力の際には注意を要する.

(1)　置換

置き換えたい文字にカーソルを合わせ文字を入力する。

(2)挿入

翻訳エディタ同様挿入キーを用いる。挿入スペースがなくなった場合には、¬Sコマンドを用いる。挿入したい位置に¬Sコマンドを入力すると、行が分割されて、３行の空行が作られるので、そこへ入力する。分割した行を結合するには、¬Jコマンドを用いる。

(3)　削除

削除キーを用いる。辞書内容全体をMTF-VSAMから削除する場合は、コマンド入力域にDELETEコマンドを入力する。

(4)　セーブ

新たに作成または修正した辞書内容は、SAVEコマンドにより保存される。MANAGEMENT FILEに履歴が残り、MTF-VSAMと更新データファイルに辞書の内容が格納される.

5．5　セーブデータのチェック

SAVEコマンドが指定されると、セーブする対象となる辞書内容について次のチェックを行う。

(1)　右カッコと左カッコの数

(2)　ストリング中はJEFコードのみ

(3)　各辞書に固有な項目の階層チェック

(3)については現在、図9の解析辞書についてのみ行っている。

項目の階層チェックを厳密に行うことは、利用者の入力ミスを防ぐ上で有効であるが、半面、辞書項目の変更・追加に対しては対応できなくなる。言語処理システム・辞書システムともに開発途中にあるため、多少の項目の変更・追加は十分考えられる。そのための柔軟性を設ける意味で、SAVEコマンドのオペランドにNOCHECKを指定できるようにした。このオペランドを指定すると(1)のチェックだけを行い、他は行わない。

6　入出力テキスト変換

入出力テキスト変換は、ファイルの世界における翻訳システムへの出入り口といえる部分である。入力変換メニューと出力変換メニューがあり、それぞれの画面から処理を行う。

6．1　入力変換

入力変換は、原文ファイルから一文を切り出し、言語処理システムの要求する形で、翻訳ファイルの原文部に原文を格納する機能を果たす。ファイルとしては日本語エディタ（PFDおよびODM）で作成した文章、情報検索システム（FAIRS-I）およびJICST抄録データの4種を考えている。

(1)　文の切り出し

一文の切り出しは、翻訳の第一歩である。日本語文はほとんどの場合、句点が一文の区切りであるが、文章によっては他の記号で区切りたいこともある。そこで、利用者がメニュー上で文区切り記号を指定し、入力変換では指定に従って一文を切り出すことにした。

(2)　翻訳するか否かの判断

本翻訳システムの利用対象の大きな柱として抄録が上げられている。抄録は翻訳対象となる表題・抄録の他に種々の書誌的事項を含む。これらは、翻訳をしたとしてもそのままの文字列が出力されるだろうが、言語処理システムに負荷をかけないために翻訳を行いたくない。そのため、入力変換では抄録データに限って、項目ごとに翻訳するしないの判断を行い、各文にその旨の情報（翻訳フラグ）を付加することにした。翻訳エディタでは翻訳コマンドが入力されると、フラグがONの文だけを言語処理システムに渡す。

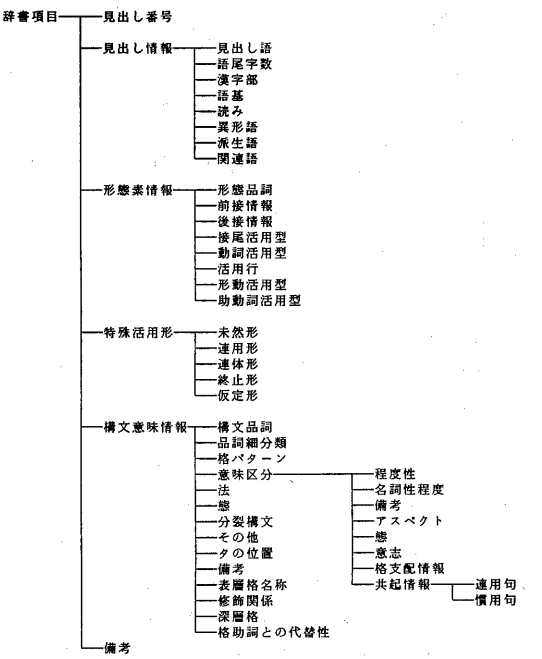

図9．解析辞書項目の階層関係

JICSTデータでは表題と抄録文の翻訳フラグをONにする。一方、FAIRS-I抄録データでは格納されている情報が各ファイルごとに違うので、利用者に翻訳したい項目を指定してもらい、入力変換は指定された項目のみの翻訳フラグを立てる。

ところで、上のように翻訳しないと判断された書誌的項目でも、訳文には組み込みたい。しかし、言語処理システムを通らないので訳文は出ない。この点を解決するため、入力変換は翻訳しない（翻訳フラグをOFFとをした）文については、同時に翻訳ファイルの訳文部にも同じ文を書き込み、原文と同じ情報がすべて伝わるようにした。

(3)　制御コード

日本語エディタにより作成された文章にはフォーマッティングのための制御コードが含まれる。制御コードは本来の文章ではないが、中には翻訳の際有益な情報を与えてくれるものがある。例えば、アン

ダーラインや段落がえ等は将来意味解析が進んでいった場合有効な情報の一つとなる。そのため現在、入力変換では制御コードを取り除くことはしていない。

6．2　出力変換

出力変換は翻訳ファイルの訳文部を読み、指定された訳文ファイルへ出力する。また翻訳ファイルの対訳をプリンタに出力する。訳文ファイルとしては、日本語エデイタ(ODM･PFD)ファイル・英文清書システム(ATF)ファイルの３種を想定している。

(1)　文字コード

言語処理システムから出力される英文は日本語(JEFコード）である。そのため訳文ファイルがATFファイルの場合は、EBCDICコードに変換する。

(2)　制御コードの処理

6.1で述べたように原文に付いている制御コードは、そのままの形で訳文に反映する。現在、訳文ファイルに出力する際、その形式に応じて制御コードを次のように処理している。

① ODMファイル　：そのまま残す
② ATFファイル　：一部対応するATFの制御コードに変換、他は削除
③ PFDファイル　：削除

7　制御・管理・その他

7．1　会話制御

総合システムは図２に示した初期メニューから利用者と対話することにより、すべての機能を実行できる。またすべての画面には、使用方法を書いた画面が用意されており、HELPキーの押下でいつでも参照できる

7．2　UTILISP-PL/I間の制御

言語処理システムおよび辞書システムはUTILISPで書かれている。一方、総合システムはプログラム開発効率および運用・管理の面を考慮し、PL/Iで開発している。よって、両言語間の制御機構が必要になる。

(1)　方式

翻訳や辞書作成のタイミングで総合システムが言語処理システム･辞書システムを外部プログラムとして呼び出す形が自然であろうが、その度ごとに言語処理システム･辞書システムの初期化を行うことになり、対話システムとして好ましくない。そこでUTILISPと総合システムをそれぞれ別タスクで起動し、同期を取りながら並列的に処理を行う。

タスク間の連絡を取るために同期処理ルーチンを用意する。これは、他タスクに事象の完了を知らせ(POST)、自タスクは他タスクの事象の完了を待つ(WAIT)という処理を行うルーチンである。

(2)　制御の流れ

図１０に制御の流れを概念的に示す。

① 主制御ルーチンが総合システムおよびUTILISPを起動する。
② UTILISPは翻訳・辞書作成関数の指定を待つタイミングで、同期処理ルーチンを実行する。
③ 総合システムは利用者と対話しながら処理を進める。翻訳・辞書作成を行いたい時に同期処理ルーチンを実行し、UTILISPの待ち状態を解除する。
④ 翻訳または辞書作成関数の処理が終了した時点で再び同期処理ルーチンを実行する。
③、④を繰り返す。
⑤ 終了キーの押下で総合システムは終了するが、この際同時に主制御はUTILISPも終了させる。

7．3　一括翻訳

ファイル単位に一度に多量の文章を翻訳する場合に、一括翻訳機能を用いる。入力変換を終了し、翻訳ファイルに格納された原文の翻訳をバッチ処理で行う。

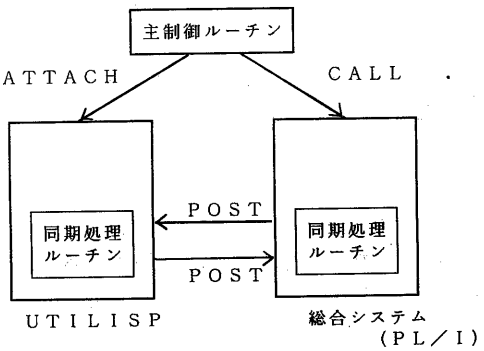

図１０　UTILISP-PLI/I 間の制御

8　あとがき

Ｍｕプロジェクトは昭和５７年度から４か年の計画である。今年度末に日英翻訳システムの総合翻訳システム実験を行う予定であり、来年度には英日翻訳システムについての実験に入る予定である。

現在、本総合システムは日英翻訳システム用として開発を行っている。同時に、翻訳の核となる部分である言語処理システム・辞書システムの開発も進められており、これらシステムの開発途上における改良・拡充に対して各システム間で正確なインターフェースをとる必要がある。

また、総合システムの本来の目的である利用者にとって使い勝手が良く、かつ効率的なシステムとするため、今後の課題として以下のような機能の拡充が求められる。

(1)　テキスト処理の拡充

入力における種々の外部提供のテキストに対するコード、特殊記号類、式・表・グラフ・図、フォーマットおよび制御コード処理と出力における清書処理

(2)　対話による翻訳

現在、利用者に対して翻訳の解析・変換・生成は一括で行っているが、翻訳の個々の段階で必要に応じ、多重画面制御による処理状況の表示を行い、翻訳手順、テキスト、辞書、文法をインターラクティブに変更可能とする。

(3)　辞書の編集と管理

辞書内容の表示をMTフォーマットのようなLISP形式でなく、表形式の一般的な表示とし、その表上での編集と内容のチェックが行える。また基本辞書、ユーザ辞書、専門辞書等が利用者やテキストにより自由に選択でき、連結して利用できる運用形式

(4)　翻訳統計の管理

・テキスト中の文字・単語・文の出現頻度
・辞書検索で利用された辞書項目の統計
・文法規則の利用統計
・翻訳の目安箱（苦情）

今後、実用化の段階で多くの人に利用してもらい、機能の拡充、翻訳作業の効率向上を押し進めていきたい。また、将来的には、日英、英日双方向の総合システムとして実現されるであろう。

＜参考文献＞

(1)　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、1983,7

(2)　工業技術院計画課・電子技術総合研究所：「日英科学技術文献の速報システムに関する研究、日本語解析辞書システム説明書」、1984,1

(3)　日本科学技術情報センター・電子技術総合研究所・京都大学：「日英科学技術文献の速報システムに関する研究、日英科学技術用語辞書データの開発に関する報告書」、1984,3

(4)　電子技術総合研究所・京都大学：「日英科学技術文献の速報システムに関する研究、言語処理システムの開発に関する報告書」、1984,3

(5)　中村順一・辻井潤一・長尾真・坂本義行・佐藤雅之：「Muプロジェクトにおける辞書の運用方式」、情報処理学会自然言語処理技術シンポジウム論文集、1984,11