技術論文表題の英和自動翻訳の試み

長尾　真，辻井潤一，建部周二（京大・工）

1. はじめに

機械翻訳システムのように、非常に多くの未解決の問題を持っているシステムを作成してゆく場合には，まず現在の技術の範囲内で比較的簡単に作成できるプロトタイプのシステムを作り，そのシステムでの経験から，残された問題と，もしそれが解決したとしたら、どの程度のパフォーマンス向上が得られるかを整理し、次に解決すべき問題点を明確にしてゆくべきであると思われる。特に、工学的立場からは，あまりに詳細な言語現象についての解明を行なっても，それが全体に及ぼすパフォーマンス向上につながらない場合もあることを考えると，プロトタイプシステムの作成と、その問題点の明確化は重要な作業になる。

我々はそのようなプロトタイプシステムとして，対象を技術論文の表題にかぎった英和自動翻訳システムを作成した。近年のデータベースシステムの発展と情報検索への要求の高まりによって，論文表題の翻訳への要求は非常に高まっており，日本では日本科学情報センター，外国では米，独，その他の国で具体的にこの種の翻訳への要求がある。

本報告は完全な形で稼動している機械翻訳システムについての報告というよりも，現在我々の作ったプロトタイプシステムで明らかになった将来の問題点，次に我々が行なうべき課題について述べることにする。

2. 機械翻訳システムの概要

自動翻訳は一般に次のようなステップをへる。

(i) 原文入力　(ii) 原文の形態素解析　(iii) 原文の構造解析
(iv) 原文の意味解析　(v) 原文情報から訳文情報への変換　(vi) 訳文の構造合成
(vii) 訳文の形態素合成　(viii) 出力プリント

これらのうち(ii)の形態素解析　(vii)の訳文の形態素合成の部分は、過去にしっかりした研究もあり，やればほぼ出来る範囲のものである。ここでは研究の焦点を(iii)と(v)にあて，(iv)についてある程度の処理を行なうという立場をとった。その理由は次の通りである。

(i) 意味の問題にはいくつかのアプローチがあるが，せまく限られた明確な世界について適用はできても，一般の分野に適用できるような確立された手法はいまだにない。

(ii) 意味を十分にあつかうためには，現実世界に関する知識も採用しなければならない。

(iii) 技術論文の表題の翻訳というような実際的な分野の場合に，(i)(ii)のような意味を積極的にあつかうのでなく，逆にどこまで意味のことを考えずに実用的なところまでゆけるかを試みる方が実際的である。

(iv) 計算機はぼう大なデータの処理にむいているから，複雑でこった方法を考えるよりも，種々の場合をなるべく多く集めて，それによって表題の解析と翻訳を考える方がよいのではないか。

このように，論文表題の翻訳を単純なプログラムとできるだけ多くの場合をお

語では、複数個の単語で1つの概念を表わすことが多い。この場合，訳出の過程で個々の単語の訳語から全体の訳語を合成していたのでは，ほとんど理解不可能な訳文が生じる。人間の専門家が専門用語の訳語を決定する場合に，個々の英語の単語の訳から合成するのではなく，その句全体が表現している概念を自然な日本語で表現することによって行なっていることを考えると，すなわち，専門用語の日本語訳を決定する際に人間の創意が入っている以上，この過程を計算機に行なわせることはほとんど不可能であり，これらの句も1つの単語と考える必要がある。そこで我々はこれらの句に相当する専門用語も1つの独立した辞書項目として設定することにした。実際には，その句中に出ている単語の辞書記述中にそれらの句を登録している。

この他，各分野でよく使われる固有の表現（Idiom的な表現）—— time varying（時間とともに変化する），settling time（整定時間）—— あるいは，英語全体にみられるidiom的表現（based on —，perpendicular to — 等の群前置詞）もまた、辞書中に記述され，辞書引きの段階で単語と同じように取扱われる。これを(2)のTLに対してRigid Idiom（以下RI）と呼ぶ。

一般に，単語の単位を大きくとればとるほど，辞書の規模は大きくなるが，逆に後での処理が簡単になり，訳文の質も向上する。特に，ある特定の専門分野に対象を固定した場合には，この手法は有効であろう。このRigid Idiomを選定する過程で，学会発行の学術用語集等を参考にしたが，用語数が少なく，ほとんど役に立たなかった。現実的な機械翻訳システム作成のためには，現在使われている用語を網羅的に収集したTerminology Data Bank的なものを用意する必要がある。

(4)表題文においては、連続した名詞，形容詞によって名詞句を構成する電文調の表現が多くあらわれる。

通常の文体では，前置詞等を使って結合される名詞が，表題文においては，単に連続するだけで結合され名詞句を構成することが多い（例 —— High Resolution Bragg Reflection Method, Pulse Compression Distance Measuring Equipment System etc.）.

これらの名詞間，形容詞－名詞間の結合関係を同定し，訳文合成の際に適切な語順変換と，助詞の挿入を行なうことはもちろん理想的ではあるが，これを行なうためには、かなり深い意味的な解析を必要とし，現在の時点ではほとんど不可能である。

一方，日本語においても，表題文においては，かなり長い漢字連続で英語と同様な表現をとることが可能であり（上記例の場合には，「高分解ブラッグ反射手法」，「パルス圧縮距離測定装置システム」となる），しかもその中での英語と日本語の語順はほとんど変化しない。

したがって，現在の我々のシステムにおいては，このような長い名詞連続内での単語間の詳細な係り受けの関係は同定せず，単にそのような長い名詞句があったということだけを簡単なTransition Network（以下TN）文法で認識することにした。現在のシステムでは日本語と英語の語順が変化するのは，前置詞，—— ed形，—— ing形等の単語があらわれた場合だけであると考え、形容詞を含む名詞連続では語順がかわらないと考えている。したがって，ここで使われるTNはこのような語順の切れ目を示す語があらわれるまでを，読みとばして1

おったデータ，法則とで行なってみようというところに研究の焦点をあてた。したがって入力文，合成文ともに形態素の処理は行なっていず，構文と変換の部分のみのテスト結果がわかるようになっている。

また、上のような翻訳のメインの流れの他に、我々が今回作成したような実験的なシステムにおいては，辞書記述やルールの変更を含む様々な試みを，できるだけ容易に行なえることが重要になる。したがって，表題文のKWIC，辞書・ルール記述のEditor等の機械翻訳システムのためのサポート・システムも同時に作成した。全体のシステム構成を図1に示す。なお，翻訳のための論文表題は，日本科学技術センターの文献速報磁気テープからとり，1万件の英文表題と，それに対して人間によってつけられた和文表題とがつけられている。辞書としては、1万件の英文表題にあらわれた9,800語（形態素処理は行なっていないので，morphological variantもすべて独立した1語として登録されている）について，品詞，日本語訳がつけられている。

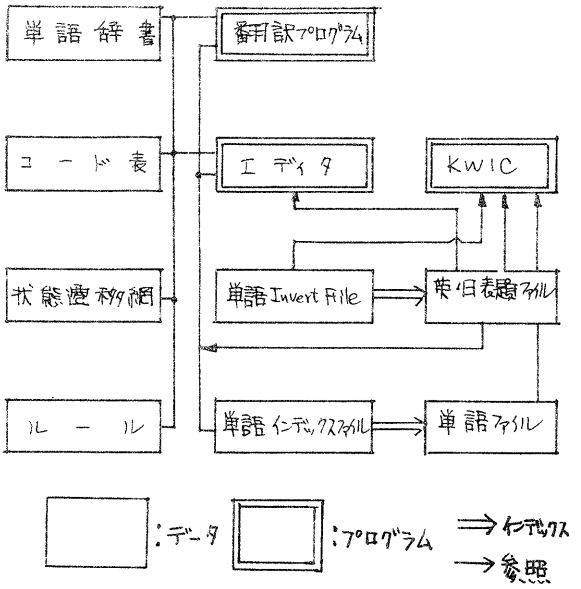


図1. システム構成

3.機械翻訳システムの構成

3.-1.システム構成の考え方

論文表題は，機械翻訳システムの対象としてみると，次のような特徴を持っている。

(1)表題文中の単語は専門用語がほとんどで、他の分野に比べて多義語の問題を排けることができる。

もちろん、専門用語であっても各専門分野ごとに別個の訳語がつけられている場合もあるが，各訳語ごとに分野コードを指定する等、簡単な措置で多義語を排けることができる。ただし，訳語と分野コードの指定は現在のシステムでは行なっていない。

(2)表題文固有の文体構造がある。

特に、Experiment on ——，Application of —— to —— 等のように，表題文全体の構造を決める英語における特有の言い回し方は、それに対応する日本語訳のパターンを用意することにより，前置詞の訳出方法など，従来困難と考えられていた問題をある程度排けるのに有効である。我々は、この種の特定な単語に固有な文体パターンをFlexible Idiom（以下FI）と呼び，各単語の辞書に登録している。

(3)専門用語は，必ずしも1語単位ではなく，句に対応することが多い。

例えば，Large Scale Integrated Circuit（大規模集積回路）のように，専門用

つの名詞句としてまとめてゆく。(ただし，{冠詞 / 形容詞} + {—ed形 / —ing形} のように明らかに形容詞的に使われていることがわかる —ing形，—ed形の場合には，語順を変換する必要がないので，TNによってそのまま読みとばされる)。
(5)表題文にあらわれる文構造は、書き換え規則で処理する必要があるほど多様ではない。

言語分析において，書き換え規則に代表される規則の体系を使用する大きな理由は，言語表現の持つ「無限」の可能性を，有限個の規則で処理したいためである。しかしながら，表題文においては，表題文全体の長さが制限されていること，埋込み文等の複雑な文体がほとんどあらわれないこと等の理由によって，すくなくとも品詞並びに還元してみる限り，規則で処理しなければならないほどの無限の可能性があるとは考えられない。しかも，前述の(3)の項で述べたように，句で表現される専門用語を1つの辞書項目と考え，また，(4)で述べたように長い名詞連続をTNによって1つの名詞に縮退する操作を行なうと，可能な品詞の並びはさらに局限される。

このような考え方から，我々は表題文の品詞列を処理する際に書き換え規則等の規則による処理は行なわずに，表題文にあらわれる可能な品詞並びをそのまま列挙しておき，これと入力表題文の品詞並びとのパターンマッチングを行なうことにより，表題文全体の訳出語順を決定することにした。入力表題文から，種々の操作を経て得られた縮退した品詞並びを，その入力表題文の骨格パターン，システムによってあらかじめ用意されている品詞並びを文体パターンと呼ぶ。したがって，入力の骨格パターンとシステムに登録されている文体パターンのパターン・マッチングによって，訳出語順が決定される。

一般に規則の数を少なくするためには，局所的な言語の構造を処理するための規則を用意し，その組み合せで文の大域的な構造を把握することになる。同じように，解析の結果を訳出する場合にも，逆方向の規則によって，局所的な構造を一部分ずつ線形表現に変換してゆくことになり，全体として訳のバランスや，より大域的な観点から決めるべき語順の決定において困難がある。

以上のような理由からも，表題文のように文体に強い制限がある場合には，我々のとった文型パターンによる訳出がすぐれていると思われる。もちろん，後述する文型パターンに付随した意味処理の指定等が，本来的には1つの統語規則について宣言されるべきものが，同じ部分構造を持ついくつもの文型パターンについて宣言されなければならないという欠点がある。

3－2．システムの構成

表題文の翻訳は次の各ステップによって行なわれる(図2に処理の流れと実例を示す —— 図2は末尾)。

Step 1：辞書引きおよびRIの処理
Step 2：LOCAL ANDの処理
Step 3：名詞連続の処理
Step 4：骨格パターンと文型パターンとのパターン・マッチング処理
Step 5：簡単な意味処理とRIを使った訳出語順の決定
Step 6：日本語文の合成

以下の節では表題文の各ステップの処理の概要について述べる。

3－2－1. Step1 ― 辞書引きおよびRigid Idiomの処理

現在のシステムでは、英語の形態素処理を全く行なわないために、morphological variantもすべて辞書項目として含まれている。もちろん、実用化の際には規則化可能な部分を規則化することによって、辞書項目数を減少させる必要がある。

ただし、同じ'V + ing'の形式で作られた語であっても、accounting, bonding, scattering, engineering等はほとんどいつも名詞的に、superconductingは形容詞的に、usingはもっぱら目的語をとってその直前の名詞を修飾して前置詞的に、determiningは他動詞的に使われる。一般に'V + ing'形はこれらのすべての使われ方をする可能性があるが、これらの語については、分野を固定するとほとんどある一定の使われ方をする。したがって、辞書記述の中にこの固定的な使われ方をするということを記述しておくことにより、後の段階での複雑な処理を排けることができる。また、time varying, based onのように、RIに組込まれているものもある。英語の形態素処理を一般的なルールで行なうことが意味をもつためには、それに対応する日本語の形態素発生がうまくルール化できて、しかも、英語の解析、語順決定等がその形態素で変形をうけた元の単語とは無関係に実行できることが必要であるが、上述のようにもとの動詞が何であったかによって、'V + ing'形の果す役割がかわる場合には、結局単語ごとにing形の付いた場合の記述が必要となり、個々のV + ing形を辞書項目として立てることと、ほとんど大差がないことになり、いたずらに処理を複雑にするだけで、利点がないことになる。

3－2－2. Step2 ― LOCAL ANDの処理

ここでは、andだけでなく、and・or・butのような等位接続詞によって作られる並列句一般の処理を行なう。実際には、対象とした表題文中には、等位接続詞としてはandだけしかあらわれなかった（1万表題中、andを含む表題722件、内2つ以上のandを含むもの51件）。

等位接続詞を含む句の解釈には、よく知られているように、scope ambiguityがある。例えば、*Controlability and Stability of the system* を「システムの可制御性と安定性」と訳すべきか、「可制御性とシステムの安定性」と訳すべきかは、結局Controlability, Stability, Systemの3語間の意味的関係によって決定せざるを得ず、意味処理の問題と密接に関連する。これを「システムの安定性と可制御性」のように、並列されている句の順序を入れ替えて、日本語においても同じ曖昧さを持つ表現に置き換えることもできるが、本質的な解決にはならない。現在の我々のシステムでは、このような曖昧さを持つ等位接続詞の解析は未解決な問題として残されている。

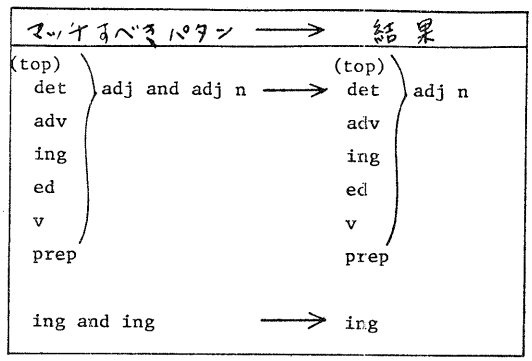


図3. 局所的に処理可能なAND

我々のシステムでは、文全体の大域的な構造を参照しなくても、局所的な品詞列だけから決定できる「形容詞 and 形容詞」等の連続だけを処理する（図3）。

この種の局所的な修飾語のand結合として処理できる用例は，722件中約10%の75件であった。それ以外のandの用例はすべて骨格パターン中に残されることになる。

実際には，骨格パターン中にandを残すというこの方法は，次の2つの理由から望ましいことではない。

(1) システムが用意する文型パターンの数が組み合せ的に増大する。

例えば，文型パターンとして'n + prep + n'があれば，これに対応して，'n + and + n + prep + n'・・・等の各名詞句がandを含む場合の文型パターンをすべて用意する必要があり，冗長度の高いパターンの数が飛躍的に増大してしまう。

(2) 次に述べるStep3では，長い名詞句を1つの名詞に縮約する操作を行なうが，次のような例のためにこの操作をandの周辺では全く実行できないことになり，骨格パターンに残ってしまう単語が増大する。

例：Combat vehicle and aircraft stabilization system

(この例では，Combatがvehicle，aircraftの両者を形容し，また動詞stabilizeの名詞化stabilizationがvehicle，aircraftの両者を目的語としてとっているために，'combat vehicle → vehicle' 'aircraft stabilization system → system'というようにandで結合されている2つの部分を単独に縮約することができない。)

以上のようなことから，andのような等位接続詞を含む句の処理は，次のStep3，4へ行くまでに処理しなければならないが，この部分の処理はかなりheuristicな手法で行なわざるを得ないと思われる。

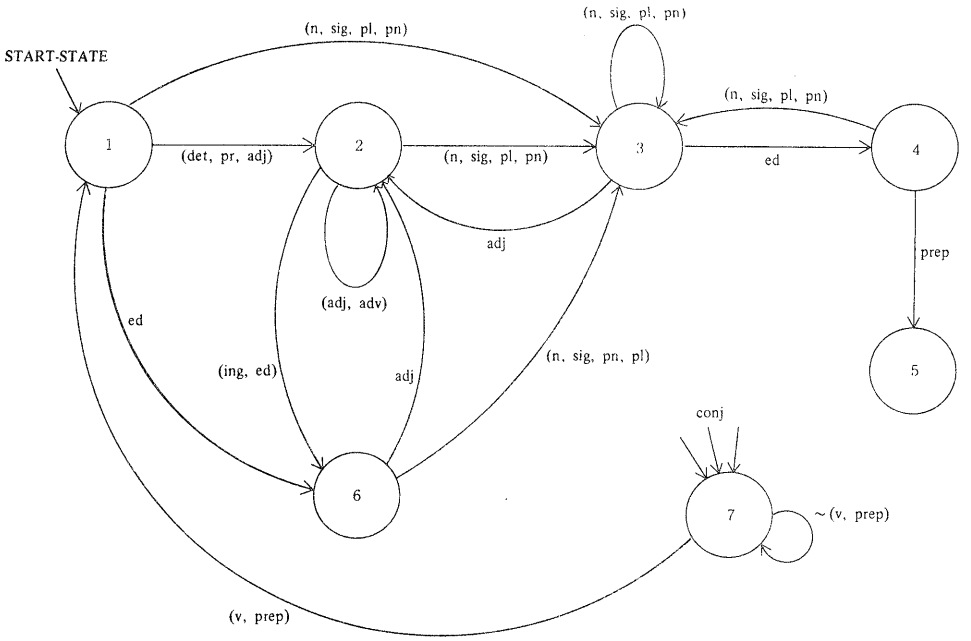


図4. 連続名詞，形容詞を処理するTN

3－2－3. Step 3 ―― 名詞連続のTNによる処理

ここまでのStepでつくられた品詞列は，RIと簡単な構造の並列句を処理しただけで，入力表題の品詞列とほぼ同じものである。従って，このままでは入力表題の数だけ異なった品詞列があり得る。このStep 3では，語順の変換を必要としない。'n + n' やそれに形容詞がそう入されたような品詞列を，1つの名詞句に縮退する。これを行なうためのTNを図4に示す。前に注意したように，このTNはあくまでAccepter型のTNで，入力品詞列に対する構造を作り出すtransducerではない。

このStepによって，次のような句が1つの単語に縮退される。

(例) This film waveguides ⟶ waveguides
High current arc discharge ⟶ discharge
Pressure control system ⟶ system
High quality phosphor creens ⟶ screens
Transmission Line Power Flows ⟶ Flows
An Automated General Purpose Test system ⟶ system
Eddy current method ⟶ method
Time varying strong magnetic fields ⟶ fields
The refractive index profile ⟶ profile
stochastically disturbed industrial Plants ⟶ Plants

3－2－4. Step 4 ―― 骨格パターンと文型パターンとのパターン・マッチング処理

以上のStepで，入力表題文に対応して，図5に示すような骨格パターンが抽出される。Step 4では，これに対してシステムに用意された文型パターンとのマッチングが行なわれる。

(1) Experimental Method of Measuring the Anisotropy Field in Small Samples
Skelton Pattern : Method of Measuring Field in Samples
品詞列 : n prep ing n prep n

(2) A Laser Doppler Technique for Measuring Flow Velocities in High Current Arc Discharge
Skelton Pattern : Technique for Measuring Velocities in Discharge
品詞列 : n prep ing n prep n

(3) Voltage Time Integral Method for Measuring Machine Inductance
Skelton Pattern : Method for Measuring Inductance
品詞列 : n prep ing n

(4) Diffuse Field Technique for Measuring Directivity Index
Skelton Pattern : Technique for Measuring Index
品詞列 : n prep ing n

(5) Apparatus for Thermopower Measurements on Organic Conductors
Skelton Pattern : Apparatus for Measurements on Conductors
品詞列 : n prep n prep n

(6) The Design of a Pressure Control System for the Pepsios Spectrometers
Skelton Pattern : Design of System for Spectrometers
品詞列 : n prep n prep n

(7) Evaluation of High Quality Phosphor Screens for Image Tubes
Skelton Pattern : Evaluation of Screens for Tubes
品詞列 : n prep n prep n

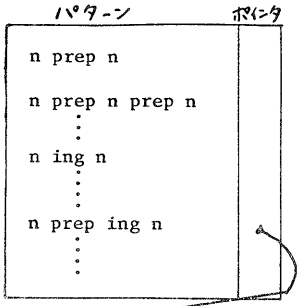
図5 入力表題文とその骨格パターン

3－2－5. Step 5 ――訳出語順の決定

Step 4で入力文の骨格パターンが見つけられるが，一般に1つの文型パターンには複数個の訳出語順があるので，このStepでは単語の意味記述とFIを用いて訳出の語順が決定される。

FIは，ある特定の単語に固有な表題文の文体を列挙しているので，これに合致する場合には，その訳出語順及び前置詞句等の訳は，

それが優先される。もし，FIを持つような単語がなかったり，もしあったとしても，入力表題文の骨格パターンがそれに合致しない場合には，システムの用意する一般の文型パターンに対する訳が採用される。図6にその模式図を示す。



	チェック関数	訳出法
1	SUBJ (3, 4)	3 4 2 1
2	OBJ (3, 4)	4 3 2 1

図6. 文型パターンと訳出語順の決定

図に示すように、各訳出語順には，その訳出語順を意味的に支持するチェック関数が用意されている。すなわち，我々のシステムにおいては，必ずしも意味的な記述が完全に与えられることを仮定していないので，Wilksの preference semantics の手法と同様に，あくまで意味処理は補助手段になっている。意味処理関数をはたらかせた結果，意味的に整合がとれていることが判った場合には，その訳出語順の優先度が上がることになるだけで，意味的な整合が確認されなかったとしても，必ずしもその訳出語順が棄却されたことにはならない。

訳出語順の決定においては —ing形の存在がもっとも大きな影響を与える。すなわち，—ing形があった場合には，measuring device，measuring temperature のように，日本語訳において他の単語との結びつき方によって語順が変化するためである。これを処理するために、我々は各動詞に関して，その動詞が主語，目的語としてどのような意味カテゴリーの名詞をとるかを記述し，これに対応して，名詞の意味カテゴリーに分類した。その1例を図7に示す。

名詞の意味カテゴリー：動詞の意味記述は，このような簡単なものではすまないが，現在どの程度の意味記述が実際に質の良い訳文を得るために必要かの検討を行なっている。

ここでいう訳出語順の決定は，実際には骨格パターン中の係受け関係 同定になっているが，この過程で問題になるのは，前置詞句の修飾先である。英語では，前置詞句はそれが修飾する語よりも後に置かれ，日本語では全くその逆になるので，英語の語順を完全に逆転させておけば，英語において前置詞句の修飾先が曖昧であれば，同じように日本語においても曖昧な表現が得られ，プロトタイプの翻訳システムとしては十分である。しかしながら，考えている英語と前置詞句が動詞（—ing形，—ed形を含む）を修飾しているか，名詞を修飾しているかによって，対応する日本語 の修飾句の語尾変化がかわる（連用修飾句か連体修飾句か）ため，このStep 5

道具	probe, instrument, equipment system, set, machine unit
理論	technique, principle, menthod approach
観点	velocity, displacement acceleration, position, energy inductance, registance capacitance, conductance temperature mobility, permittivity intensity level parameter coefficient, time, angle
物体	laser, beam, light solenoid, coil film, tape, disc metal MOS, device, waveguide diode, transistor oil, water, liquid, solid air, earth star, stone

図7. 名詞のカテゴリー分類

における係り受け関係の決定は、より質の高い訳文を得るために重要である。現在のシステムでは、次にくるStep 6の日本語の形態素発生の処理を行なっていないため、このStep 5で動詞にかかるとされた前置詞句もすべて連体修飾語として生成しており、これが訳出結果の質を落す原因にもなっている。このStep 5の訳出語順の決定をうまく形態素発生に結びつけるStep 6を開発する必要がある。(現在のシステムでは、なるべく動詞的概念も名詞化して表現しているため、これによる不自然さはいくぶん緩和している。——(例) A Method of measuring velocity by doppler effect

⟶ 「ドプラー効果による速度測定手法」(現システム)

⟶ 「ドプラー効果によって速度を測定する手法」)

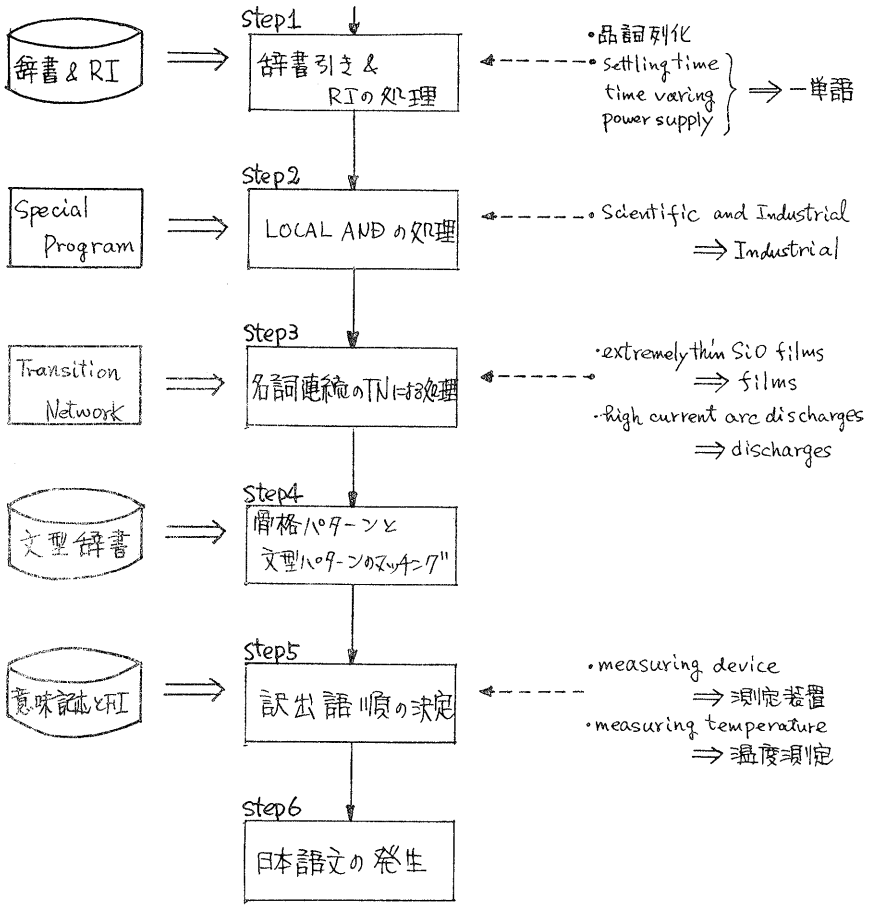
4. おわりに

現在、本システムは京大大型機センターでPL/1を使って実現されている。本システムに翻訳された実行例を図8に示す。現時点での問題点は、本文中に述べたが、これ以上により本格的なシステムを作成するためには、文型パターン、RI、FIの網羅的な収集と、単語数の増加が必須である。現在、表題数を増加させた時に、骨格パターンがどのように増えるか、JICST発行の文献テープ中にある表題をかなりの安定度で訳出するためには、どの程度の文型パターン、RI、FIを用意すれば良いか等の評価を進めている。これらの評価の他に、訳文の質を人間の翻訳と比較して、定量的に検討してゆくことも重要であろう。

今後に残されたシステム構成上の問題点は① AND の処理、②名詞、動詞の意味記述、③語順が確定した時点での日本語の形態素発生 等が考えられる。

```
▶GENERATION OF STANDARD EM FIELDS USING TEM TRANSMISSION CELLS.

 < 001 N     > < 002 PREP > < 003 N     > < 004 SIG  > < 005 PL    > < 006 PREP >
 < 007 SIG   > < 008 N    > < 009 PL    >

 < 001 N     > < 002 PREP > < 003       > < 004      > < 005 N     > < 006 PREP >
 < 007       > < 008      > < 009 N     >

SKELTON PATTERN:  GENERATION  OF  FIELDS  USING  CELLS

 BUNBANGO=     22**DAI      1 BANME**

GENERATION OF STANDARD EM FIELDS USING TEM TRANSMISSION CELLS.
ＴＥＭ伝送セルを用いた標準ＥＭ場の生成
BUN BANGO=    23
▶TRANSIENT ELECTROMAGNETIC PROPERTIES OF TWO, INFINITE, PARALLEL WIRES.

 < 001 ADJ   > < 002 ADJ  > < 003 PL    > < 004 PREP > < 005 N     > < 006 ADJ  >
 < 007 ADJ   > < 008 PL   >

 < 001       > < 002      > < 003 N     > < 004 PREP > < 005       > < 006      >
 < 007       > < 008 N    >

SKELTON PATTERN:  PROPERTIES  OF  WIRES

 BUNBANGO=     23**DAI      1 BANME**

TRANSIENT ELECTROMAGNETIC PROPERTIES OF TWO, INFINITE, PARALLEL WIRES.
二無限平行ワイヤの過度電磁性質
```

図8. 翻訳システムの実行例



翻訳ステップの例

入力 : Industrial and Scientific Techniques for Measuring field Effect Mobility

Step1 : Industrial and Scientific Techniques for Measuring Effect Mobility

Step2 : Scientific Techniques for Measuring Effect Mobility

Step3 : Techniques for Measuring Mobility

Step4 : <文型パターン> | n prep ing n |

訳出結果 : 電界効果移動度測定のための工業的&科学的手法

図2. 翻訳システムの流れと実行例