

英日機械翻訳システムSHALTにおける 単純名詞句の翻訳手法について

堤 豊 寛 義郎 堤 泰治郎

日本アイ・ビー・エム株式会社 サイエンス・インスティテュート

本稿では、IBMの計算機マニュアルを対象にした英日機械翻訳システムSHALTの特徴の1つである、単純名詞句(SNP)の翻訳手順について述べる。SHALTでは、技術文を翻訳することを目的としているため、ソフトウェア、ハードウェア、新しい技術などの名前をはじめとして、長い名詞句が頻繁に出現する。これらの名詞句を自然な日本語に翻訳することで、翻訳文の品質をかなり高めることができる。しかし、名詞句の構造は、文のレベルの構造とは、かなり異なっているため、名詞句の処理を文のレベルのトランスファー処理部から分けて行なう方が効率的であると思われる。

The Simple Noun Phrase Translation Component
of the English-to-Japanese Machine Translation System SHALT

Yutaka Tsutsumi Yoshiro Kakei Taijiro Tsutsumi

Science Institute, IBM Japan, Ltd.

We have developed the English-to-Japanese machine translation system SHALT (A System for Human-Assisted Language Translation). SHALT has several unique features. This paper describes one of the features, i.e., SNP (Simple Noun Phrase) translation.

Because technical documents include long noun phrases such as names of programs, devices or new techniques, it is worthwhile to use a noun-phrase processor independently from the transfer process of the sentence. In this paper, we report the mechanism of the SNP processor and the result of the translation test.

1. はじめに

英日機械翻訳システム SHALT (System for Human-Assisted Language Translation) は IBM の計算機マニュアルを対象とした、トランスファー方式の機械翻訳システムである。計算機マニュアルを初めとして技術文は、一般に文の構造が比較的単純である反面、ソフトウェア、ハードウェアおよび次々と生まれている新しい用語等のために長い名詞句が頻繁に使用され、それらの名詞句をうまく訳さなければ、翻訳された文は非常に読みづらくなる。また、これらの名詞句をそれぞれ辞書に登録すれば、訳語は高品質になるが、辞書は膨大となり、保守・管理が非常に困難になることが予想される。SHALT では、特別に単純名詞句の翻訳だけを行なうプロセスを設けて名詞句の処理を行なっている。これにより、長い名詞句の翻訳を効率的に、かつ、高品質に行なうことができる。本稿では、SHALT 上に実現されている単純名詞句の翻訳方式について述べる。

2章以降では、まず最初に、SHALT システムにおける単純名詞句翻訳の位置付けを述べ、次に名詞・形容詞などの訳し分けについて具体的な手法を説明し、そして最後に、翻訳結果を示し今後の検討課題を考える。

2. SHALT システムの概要と単純名詞句の位置付け

本章では、英日機械翻訳システム SHALT の概要について述べる。SHALT の翻訳プロセスを図 2. 1 に示す。各コンポーネントの詳細については参考文献 [1] [2] を参照されたい。図に示されるように、単純名詞

句の処理は、英日木構造変換の中で呼び出されるが、この単純名詞句の処理を独立させた理由を次に示す。

- 技術文では、他の分野に比べ名詞句の長さが一般に長い。
- めったに出現しない用語を辞書に登録するのは非効率的であり、また、新製品等が作られる度に、新しい用語が出現し、それらを全て辞書に登録し維持・管理するのは、困難が伴うので、可能な範囲で個々の単語から合成して、訳語を作ることが必要である。
- 一般に、単語の意味を考えない構文解析では、名詞句内の係り受けの情報が不十分で翻訳の段階でもう一度解析をやりなおす必要がある。
- 英語においても、日本語においても、名詞句内の構造は文のレベルの構造とは異なっているため、名詞句の翻訳を木構造変換部で行なうのには、無理がある。

単純名詞句処理ルーチンは英日木構造変換の中で呼び出され、入力として、英語の単純名詞句が渡され、翻訳された日本語とその意味マーカーを結果として返す。(意味マーカーについては、4章で詳しく述べる。) なお、英文解析からも単純名詞句処理ルーチンが呼ばれているが、これは英文解析の前処理として専門用語検索が英文解析の効率を上げるために、解析に先立って行なわれる。

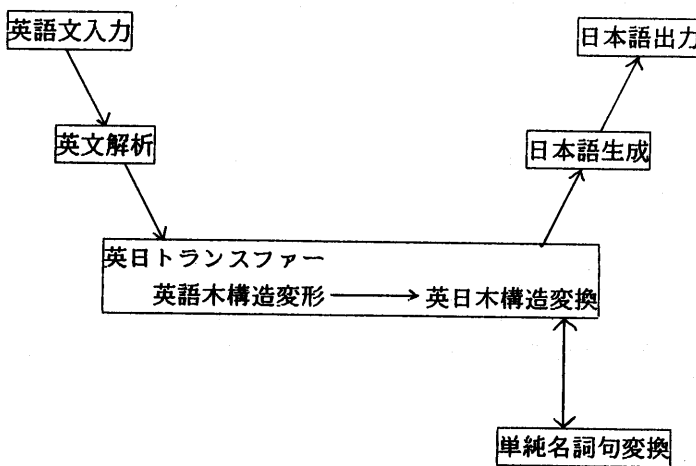


図 2. 1 SHALT 翻訳プロセス

3. 単純名詞句の定義と翻訳の方針

ここでは「単純名詞句」という用語は名詞句のうち比較的単純なものという意味で、以下にその定義を与えておく。

単純名詞句とは、名詞、形容詞、冠詞、副詞、形容詞相当語句 (a variety of 等)、およびそれらを結合する接続詞により構成される名詞句をいう。

つまり、ここで単純名詞句と言っているものは、一般にいう名詞句のうち後置修飾を除いたものといえる。以下に、単純名詞句とそうでないものの例を記す。

(1) 単純名詞句の例

processor storage
the user exit routine
IBM defined sequence
(IBM-definedは形容詞相当語句として扱われる。)
a variety of IBM licensed and non-licensed programming support
(a variety of は形容詞相当語句として扱われる。)

(2) 単純名詞句に含まれない例

new data set called SORT.SAMPJCL
program for user

単純名詞句をこのように定義して、文のレベルと切り離すことによって、英日木構造変換部の仕事を明確にできる。

ここで我々が単純名詞句用の翻訳プログラムを作成するにあたり、基本とした方針を述べる。

- できるかぎり、実際の翻訳者の訳語と同じ結果が出るようにする。基本的には、プロの翻訳者が翻訳した、IBMの計算機マニュアルでの対訳と一致するようにした。
- 辞書をなるべく小さく抑える。
- 翻訳の対象分野を計算機マニュアルに限定し、できるかぎり、その分野内での情報を有効に活用し、訳語の品質を向上させる。このために、本システムでは、意味マーカを有効に活用している。この意味マーカについては次章で説明する。

4. 意味マーカ

SHALTシステムでは、辞書項目数を減らし、効率的に翻訳を行なうために、計算機マニュアル分野における意味マーカを利用して。全ての名詞に対して、1つ以上の意味マーカが付けられている。意味マーカは、計算機マニュアル中に現れる名詞の表わす意味をクラス分けしたもので、例えば、programは、"L

E"という意味マーカを持っている。また、operatorは、"LE" (演算子) および"HM" (操作員) という、2つの意味マーカを持っている。現在、本システムでは24個の意味マーカを使用している。これらはIBMの計算機マニュアルに出現する名詞を意味的に分類したもので、かなり専門的なカテゴリーに分類されている。

LC	Logical Container
LE	Logical Entry
LP	Logical Path
DM	DocuMent
ST	STate
TH	Theory, idea and technique
FA	Feature and Ability
IF	InFormation
AT	ATtribute
VA	VA lue of ATtribute
HM	HuMan
UD	Unit or Device
WK	WorK
PS	Predicate/Simple sentence
AP	Attribute of Predicate
SL	Supply for computer system
PT	ParT
DT	Documentation Terms
ML	Material
TM	Time
PL	PLace
PN	Person's Name
PO	POint
OG	OrGanization or department

表4.1 SHALT意味マーカ一覧

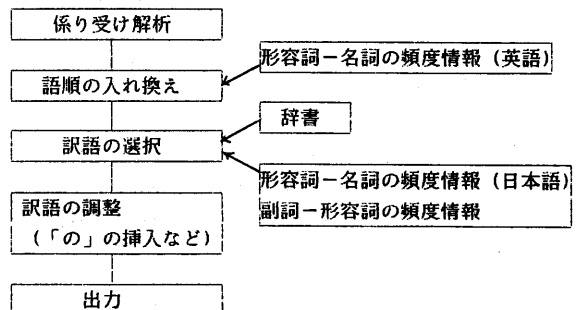


図5.1 単純名詞句の翻訳手順

5. 単純名詞句の処理手順

単純名詞句の翻訳の手順を図5. 1に示す。長い名詞句の中には、ソフトウェアの名称など、そのまま辞書に登録する以外に処理できないものが多々ある。これらの技術用語の検索は、単純名詞句処理部内ではなく、英文解析の前処理として行なわれる。長い技術用語をあらかじめ認識することにより英文解析の効率を上げ、かつ解析の誤りを減らすことができる。ここで用語辞書とは、計算機関係の技術用語を約10,000語収録したIBM情報処理用語英和対訳集[5]の機械可読形式のものを基礎にしている。ただしここでは、そのうち2語以上からなる用語についてのみ使用している。

図5. 1中で英語の共起頻度の情報は、具体的には形容詞と名詞の英文中での係り受け頻度の情報である。これは単純名詞句内での係り先を決定するときに使用される。また、訳語の共起頻度情報は、名詞一名詞、形容詞一名詞、副詞一名詞、副詞一名詞があり、それぞれ訳語を決定するときに使用される。そのほかにも、いくつかのデータが使用されているが、詳細は後続の章で述べる。

6. 係り受けの解析

係り受けの解析は、本システムのように、単純名詞句をプログラムによって処理するシステムでは、非常に重要である。なぜならば、形容詞の係り受けや接続詞などかなり意味的な処理を必要とするものがあり、これらは構文的な解釈のみでは、係り受け関係を決定できないからである。

単純名詞句内の2項間の係り受け関係には、名詞一名詞、形容詞一名詞、副詞一名詞がある。以下それぞれについて説明する。

まず、名詞一名詞についてであるが、英語の名詞連続の場合、用語辞書にうまく登録すれば、95%程度が左分かれ木になる。[3]よって、名詞連続については、一括して左分かれ木として取り扱うことにする。

次に、形容詞一名詞の関係では、係り受けの関係が、大きく訳語を左右する。例えば、

normal program termination

では、normal は、program ではなく termination に係るので、訳語としては、「プログラムの正常な停止」にすべきである。このため、本システムでは、英語の形容詞と名詞の共起頻度の情報を使用して係り受けを決定している。

副詞一名詞の場合は、殆どの副詞が直後の形容詞を修飾するので、特別な処理は必要としない。

7. 訳語の選択

係り受けの関係が得られたら、これに基づいて2項関係を抽出し訳語の選択を行なう。以下、2項関係に基づく単語の訳し分けについて述べる。

7. 1 名詞連続の処理

名詞連続は、一般に多くの場合、技術用語である。これらを総て用語辞書に登録すれば翻訳の結果は、非常に良くなるが用語辞書は膨大になり、かつ新しい用語に対処できない。そこで本システムでは、用語辞書への登録をかなり制限し、各単語の訳語を合成するという手法をとっている。この場合、名詞連続の処理では次のような問題がある。

7. 1. 1 名詞のあいまいさの解消

技術文書に現れる名詞は、一般にあいまいさが少ないと思われるが、それでも5%強の語が2つ以上の訳語を持っている。主名詞については、単純名詞句の範囲内では、多義性の解消は、不可能であるが、その他の名詞については、主名詞との関係から、殆ど訳語を決定することができる。その方法として、次のようなデータを用いる。

- ・ 日本語名詞の2項関係の頻度情報
- ・ 名詞の意味を抽象化した意味マーカールを利用した2項関係の頻度情報

現在、SHALTでは、4章で述べた24種類の意味マーカールを使用しているが、意味マーカールを使用する利点を次に挙げておく。

- ・ シソーラス等を使用した場合に比べて辞書の規模が小さくてすみ、保守・管理が容易である。
- ・ 辞書の更新の際にも、単語の持つ意味を考えて意味マーカールをつけるだけでよいので、簡単である。
- ・ 頻度情報などを表として蓄える際に、各単語をそのまま記憶するのに比べ記憶域が小さくてすみ。

これにより、実際に訳語を名詞一名詞の関係から決定する場合には、

- ・ 訳語一名詞
- ・ 訳語一名義マーカール
- ・ 意味マーカール訳語
- ・ 意味マーカール名義マーカール

の4つの頻度情報をもとに訳語を選択する。実際の計算機マニュアルからあいまいさのある語を含む名詞連続を任意に約100個抽出して翻訳した結果、約87%がこの方法によりうまく訳せることが確認できた。

7. 1. 2 「の」の挿入

長い名詞連続を日本語らしく訳す場合には、「の」の挿入は不可欠である。例えば、「PL/I language feature

s」という名詞句に対し、「PL/I言語特徴」では、意味の把握が容易ではないが、「PL/I言語の特徴」といえば、すぐに判る。このように、「の」の果たす役割は極めて大きい。「の」の挿入については、既に基本的な研究がされていて、簡単に述べると次のようになる。

- 1) 動作を表す名詞の前には、「の」を挿入する。
- 2) 動作名詞以外の名詞と属性を表す名詞の間には、「の」を挿入する。
- 3) 1、2以外で、動作名詞と属性名詞の間には、「の」をつけない。
- 4) 長さを考慮して、「の」を挿入する。

本システムでは、意味マーカを使用して、これらの条件をもう少し細かく設定しており、現在約20の規則を持っている。以下にその主なものを列記する。

- 5) "HM"「人間」と"IF"「情報」の間には、「の」を挿入する。
- 6) "DM"「ドキュメント」と"LE"の間には、「の」を入れない。
- 7) "WK"と"TH", "AT", "IF", "UD", "LE", "FA"の間には、「の」を入れない。
- 8) 漢字かな混りの語の前には、「の」を挿入する

SHALTでは、文の翻訳のほかに独立した名詞句(箇条書き、章の名前など)の翻訳も行なっているが、一般にこのような名詞句では、「の」の出現頻度が小さい。このため、SHALTの単純名詞句処理でも独立した名詞句の翻訳かどうかによって「の」を入れる基準が異なっている。

7.1.3 接辞化

name, area, length等は、単独で用いられる時は、それぞれ、「名前」、「区域」、「長さ」と訳されるが、名詞連続中の主名詞として現れるときは、「名」、「域」、「長」のように、接尾語として訳すのが望ましい。しかし、word lengthのように、前にくる名詞が、漢字1字で表される言葉の場合には、「語の長さ」と訳す。接辞化と「の」の挿入のアルゴリズムを図7.1に示す。

7.1.4 省略

"processor unit"は、「処理装置」と訳されるべきであるがprocessorは、単独で用いられても、「処理装置」という訳語を持っている。このため、processor unitをそのまま訳すと、「処理装置装置」と重複した訳がでてしまう。大規模な日本語名詞の包含関係の情報を用いれば、修正を行なえるが、ここでは、できるだけ少ない情報で簡単に処理をするため意味マーカを使っている。

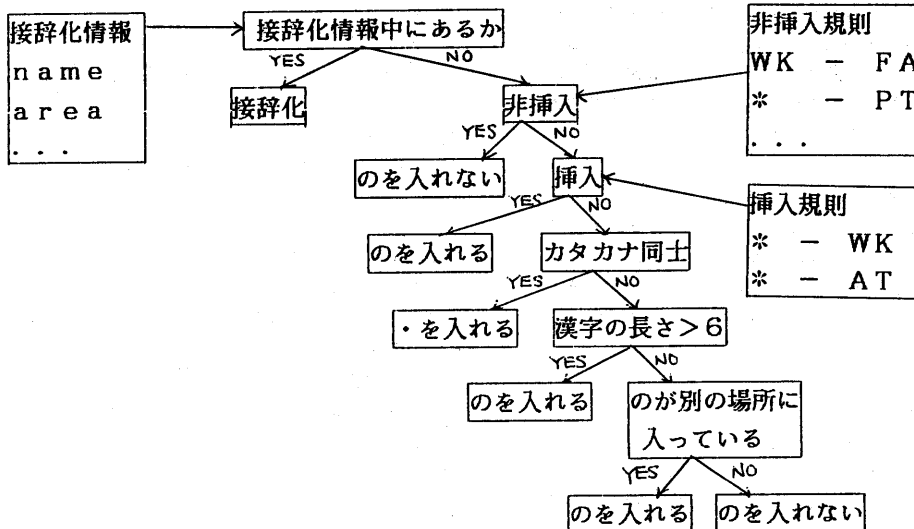


図7.1 接辞化と「の」の挿入アルゴリズム

- ・ 名詞連続があって、それぞれの訳語のうち重複したものがあつ、かつ、それぞれの意味マーカーが等しい場合には、重複部分を削除する。

この方法を用いれば、processor[UD], unit[UD]では、重複部分の削除を行なうが、international youth[HM], year[AT]では、意味マーカーが異なるため、重複部削除は行なわれず、「国際青年年」が訳語となる。

7.2 形容詞の処理

形容詞の翻訳では、主に次のような処理を考えなければならぬ。

- ・ あいまいさ
- ・ 訳語の語順

次に各々について述べる。

7.2.1 あいまいさ

計算機マニュアル中で使用される形容詞には、一般に2つ以上の訳語を持つものが多いが、訳語の間には、余り意味的に大きな差はなく、ニュアンスの違いといった感じである。形容詞の訳し分けは、形容詞と、その形容詞が修飾している名詞との間の2項関係により行なう。

この場合、頻度情報として、

- ・ 形容詞(訳語) - 名詞(訳語)
 - ・ 形容詞(訳語) - 名詞の意味マーカー
- を利用して、訳語を決定している。

英語を日本語に翻訳する場合、必ずしも、英語で形容詞であったものが日本語でも形容詞にならないことがある。例えば、「logical unit」は、「論理装置」と訳される。これは、そのまま技術用語に入れることも可能ではあるが、一般に、logicalの後ろの名詞の意味マーカーがUD(Unit Device)や、LE(Logical Entity)の場合には、「論理」という訳語になるため、「論理」と「論理的な」を頻度情報として持つことによって、用語辞書の見出し語を減らすことができる。

7.2.2 訳語の語順

一部の形容詞では、係り先の名詞によって語順を変更して訳出したほうが自然な場合がある。例えば、「overall system operation」では、「全体的なシステム操作」という訳よりも、「システム操作全般」のほうが望ましい。このほか、whole, entire等も同様である。

7.3 ハイフン

英語の文では、2つ以上の単語をハイフンで結合して使われることがしばしば行なわれる。これらのハイフンを含んだ単語をすべて登録すると、膨大な辞書が必要となり、現実的とはいえない。

本システムでは、このようなハイフンを含んだ単語は、それを構成する個々の単語の意味から合成して、翻訳する方式をとっている。この方法により、自然な訳語を比

較的小型の辞書で生成することができる。また、使われている用語全てを辞書に登録する場合には、未知語に対して訳語が得られなかったが、この方法では個々の単語が登録されている限り、訳語を生成することができる。以下にその例を示す。

- ・ system-to-operator messages
「システムから操作員へのメッセージ」
- ・ 1-to-8 characters
「1～8文字」
- ・ user-written program
「ユーザーが書いたプログラム」
- ・ non-SNA system
「非SNAシステム」

7.4 副詞の翻訳

副詞の訳語は、副詞と、その副詞が修飾している形容詞との間の2項関係によって、決定される。このために、副詞-形容詞の頻度情報が使用される。単純名詞句の中で使用される副詞には多義性のあるものは少なく、また、係り受けの関係も、ほとんどの場合直後の形容詞に係るため、特別な処理は必要としない。

7.5 ラベルの翻訳

「chapter 4」の4の部分を実ラベルと呼ぶが、日本語では、ラベルの表現には、次の3種類がある。

- ・ model 3
「3型」のように、ラベルの後ろに接辞がくるもの。
- ・ table 5
「表5」のように、ラベルの前に、接辞がくるもの。
- ・ chapter 4
「第4章」のように、接辞が前と後ろにくるもの。

ラベルの翻訳は、非常に些細なことのようには思われがちであるが、これを「章4」と訳すか、「第4章」と訳すかで、読みやすさの度合がかなり違う。さらに、技術文では、ラベルの出現頻度は、他の分野に比べ非常に高く、これをていねいに訳すことは不可欠と思われる。

これらの訳し分けは、ラベルの直前にくる名詞によってある程度判別できるため、本システムでは、chapter, model, table等の名詞をグルーピングし、テーブルとして使用し、これらの訳し分けを行なっている。

7.6 接続詞の翻訳

単純名詞句中に現れる接続詞には、次のようなものがある。

- ・ 名詞と名詞を接続するもの。
- ・ 形容詞と形容詞を接続するもの。
- ・ 副詞と副詞を接続するもの。

接続詞は、表層の語順通りに翻訳しても殆ど問題がないが、より高品質の翻訳を行うためには、特に“and”について、次のような処理をすることが要求される。

7.6.1 名詞と名詞を接続する接続詞

名詞と名詞の間のandについては、単に、「と」や「および」と訳してもかまわないが、次のようなものは、訳語を分配するほうが好ましい。

- ・ A and B C において、Aの意味マーカ―とBの意味マーカ―が同じで、かつ、Cの意味マーカ―とは異なっている場合。

(例) read and write operation

読み取り操作および書き出し操作

- ・ A B and C D において、BとCの意味マーカ―が同じで、かつ、AおよびDの意味マーカ―とは異なっている場合。

(例) program development and test environment

プログラム開発の環境とプログラム試験の環境

これらは、それぞれ「読み取り操作および書き出し操作」、「プログラムの開発と試験の環境」と訳しても、間違いではないが、技術文の翻訳では、出来るかぎりあい昧さの少ない翻訳文を出力することが要求されるため、例で示したように訳語を分配して、あいまいさを減らした訳語を出力する必要がある。

7.6.2 形容詞と形容詞を接続する接続詞

形容詞間のandについては、「かつ」と「および」の2つの意味で使われるので、どちらの意味で用いられているのかを判断する必要がある。本システムにおいては、「および」の意味で用いられる形容詞の対を調査した結果をデータとして持っており、これに基づいて「かつ」か「および」かを判断している。

- ・ complete and effective computing facilities
完全で効果的な計算機能
- ・ temporary and permanent assignments
一時的な割当てと永続的な割当て

名詞間のandでは、input and output deviceを「入力と出力の装置」と訳しても誤訳ではないが、形容詞間のandでは、complete and effective facilityを「完全な機能と効果的な機能」と訳せば、誤訳である。このように、「かつ」の場合と「および」の場合では、日本語に翻訳した場合に訳語の形が大きく異なるため、この判断は非常に重要であるといえよう。

8. 単語置き換えによる翻訳と単純名詞句処理の翻訳結果の比較

本章では、単純名詞句処理の結果と、単純に単語を置き換えることにより得られる結果を比較することによって、本システムの単純名詞句処理の特徴について述べたい。図8.1に、いくつかのサンプルを示す。最初の例は、「の」の問題である。このように「の」があるかないかで大きく読みやすさが変わる。2つめの例と、3つめの例は、対照をなしておりphysicalを「物理」と訳すか、「物理的な」と訳すかの問題である。単語置き換えの翻訳では、原則として訳語は1つしか辞書に登録されないため訳語を選択することができず、どちらかが犠牲になる。最後の例は、語順の問題である。IBM suppliedは日本語になった場合に「IBMが提供する」というように叙述的に訳されるので、これを先頭に置かないと、「各種の」が「IBM」に係るように受け取られてしまう。このように、単純名詞句処理を行うことによって、きめ細かい翻訳が可能になる。

例1

PL/I language features

PL/I言語特徴

PL/I言語の特徴

例2

physical unit

物理的な装置

物理装置

例3

various physical transmission media

さまざまな物理的な伝送手段

さまざまな物理的な伝送手段

例4

a variety of IBM supplied licensed and nonlicensed programs

各種のIBM提供のライセンスおよび非ライセンス・プログラム

IBMが提供する各種のライセンス・プログラムおよび非ライセンス・プログラム

図8.1 単純名詞句処理の結果と単語置き換えの比較例

9. 単純名詞句の翻訳結果および考察

本章では、単純名詞句について、テストを行なった結果について述べる。翻訳結果の評価基準については、まだ定説がないが、ここでは一応4つの段階に分けた。

- (1) 完全に翻訳者の訳語と一致する。
- (2) 日本語の名詞句として、十分意味が認識できる。
- (3) 「の」の不足や入れ過ぎ。
- (4) 意味が違う、または結果が出ない。

以下に、IBM計算機マニュアルから、無作為に抽出した、約2,000個の2単語以上から構成される単純名詞句に対する翻訳結果を示す。

	該当数	百分率 (%)
(1)	1202	59.8
(2)	526	26.2
(3)	82	4.1
(4)	199	9.9

このうち、(1)と(2)が全く変更なしに使用できるもので86%が成功したといえる。

誤訳の約30%は、「の」の問題である。このうち、およそ半数が「の」の入れ過ぎである。全く「の」を入れないと仮定すれば、成功した86%のうちの約4分の1が意味が変わってしまうので、全く「の」を入れないよりは、かなり文の読みやすさに貢献していると言えるであろう。ただし、日本語として自然な文を生成するためには、まだまだ検討の余地がある。「の」をどこに入れるかについては、我々はIBMの計算機マニュアルを参考にしたが、現実には、複合名詞句の解析についての研究が進んでいるので、その成果も取り入れる必要があるであろう。

(4)にふくまれる199例の殆どが訳語の選択誤りである。この大半は、名詞連続であり、頻度情報の不備が大きな原因と考えられる。今後はもっと大量のデータについて頻度をとる必要がある。また、用語辞書に登録すべきものが登録されていず、これが誤訳の原因となる場合も多い。これについては、用語辞書への登録の基準を定める必要があると思われる。

以上SHALTにおける単純名詞句の処理手順について述べた。SHALTは現在まだ研究中で性能の向上を続けているが、単純名詞句の部分はほぼ飽和していると思われる。今後は、用語集も含めた辞書の拡充が本システムの性能に大きくかかわると予想される。

参考文献

- [1] 堤泰治郎他 : 英日機械翻訳システムSHALTにおける英日トランスファーについて (自然言語研究会報告53-4, 1986.1)
- [2] 原田雅弘他 : 英日機械翻訳システムSHALTにおける日本語生成 (自然言語研究会報告53-5, 1986.1)
- [3] 堤泰治郎他 : 情報処理用語の英日翻訳について (情報処理学会第27回全国大会)
- [4] 堤豊他 : 英日機械翻訳システムSHALTにおける単純名詞句の翻訳 (情報処理学会第31回全国大会)
- [5] IBM用語委員会編 : IBM情報処理用語英和对訳集 (N:GC18-8005-1 1983.3)