自 然 言 語 処 理 78−19
（1990. 7. 20）

# 英日機械翻訳における局所解析について

中瀬 純夫
（株）カテナ・リソース研究所

鄭 敏
慶応義塾大学

自然言語処理において、固有名詞表現や数や時の表現のような定型的なパターンを局所的に認定・処理することは、文全体の構文・意味的解析とは別に、実用上非常に重要な課題である。英日機械翻訳システム *Star* では、このような局所的解析・翻訳処理（局所解析、*LOCT* ）を基本アルゴリズムの一部として組み込んである。ここでは英語テキストの解析と翻訳のために、どのようなパターンの処理を行なうことが要求されるかを検討し、局所解析の具体的な実現方法について報告する。特に、WFS( *Well−Formed Substring* ) の概念を利用した一般化パージングの概念がここでは用いられている。

# Local Context Parsing in English-Japanese Machine Translation

*Sumio Nakase*

Catena-Resource Laboratories, Incorporated
27, Ichiban-chou, Chiyoda-ku, Tokyo, 102 JAPAN

*Zheng Min*

Department of Administration Engineering,
Faculty of Science and Technology,
KEIO University
3-14-1, Hiyoshi, Yokohama, Kanagawa-ken, 223 JAPAN

The recognition and processing of fixed patterns containing time, numeral, and proper name expressions is of upmost importance in the practical application of natural language processing. The English-Japanese machine translation system *STAR* contains analysis and processing of such patterns (*Local Context Parsing*) within its base algorithm. This paper examines the need for applying the process and reports on concrete ways of its realization. In particular, it explains a form of parsing which includes the concept of *Well-Formed Substring* ( WFS ).

## 1. 翻訳アルゴリズムと局所解析

### 1．1 翻訳エンジンＳｔａｒの基本アルゴリズム

本報告では、自然言語解析における「局所解析」という概念を、英日翻訳エンジンＳｔａｒにおけるその実現という観点から論ずる。そこで、Ｓｔａｒの基本アルゴリズムおよびいくつかの基本概念について準備しておく必要がある。Ｓｔａｒの解析処理の詳細については文献［１］を参照されたい。

Ｓｔａｒ英日翻訳エンジンの基本構成は概略下記の通りである：

・形態素解析

入力文中の全ての可能な「単語」を認定、それぞれに辞書項目を付与してそれらの接続構造（ＷＦＳ構造）を作成する。現在までのＳｔａｒシステムでは、本稿での主題である「局所解析」がこのフェーズに組み込まれている。

・表層構文解析

表層の「ＷＦＳ構造」に対するＣＦＧパージングを行い、そこで成功したすべての「統語木」あるいは表層解析履歴を「パック」したＡＮＤ／ＯＲグラフ表現を作成する。ここでは、ＷＦＳ構造を拡大しながらＡＮＤ／ＯＲグラフ作成を行うＷＦＳパージングというアルゴリズムを用いている。

・深い解析

表層の解析に用いたＣＦＧルールに随伴する構造作成規則により翻訳中間構造を作成、それに対し辞書記述などを用いた共起制限や中間構造自体の評価・変形などを行う。同時に、ＡＮＤ／ＯＲグラフによって表現された数多くの表層解釈の中から、コスト最小探索問題を解く形で、最適あるいはそれに近い幾つかの解釈を選択する。

・英日構造移行

翻訳中間構造は構文的な格支配の構造に共起関係などが重ね合わされた構造である。これを日本語の（表層的）構文構造に変換する。語順の入れ替え、否定や疑問などの各種情報の移動、基本的な「送り」の決定などが行われる。

・日本語文生成

日本語中間構造をもとに日本語文を作成する。解析処理などで特定された訳語を訳出すると共に、共起制限、否定や疑問をはじめとする各種情報、接続関係などを用いた、活用、助詞・助動詞などの「送り」の制御なども行う。

この他に、入力テキストの書式の処理と文単位への分解、翻訳結果の各種出力・表示形態への編集処理、翻訳処理への部分的介入や翻訳エディタ機能などを含む「ユーザ・インタフェイス」機能、そして翻訳処理のための辞書システムと文法システム、各種分析、チューンアップ・サポート機能がある。この中で、入力テキストの形態的処理は本来、本論との関係が特に深いものであることには留意されたい。また、Ｓｔａｒの特色の一つとして、辞書構成にシステム辞書とオプション辞書という２種類の形態のものを用いていることがある。後者は単純な記述形式の辞書であるが、局所解析用辞書データのためにも用いられている。

Ｓｔａｒの翻訳アルゴリズムの際だった特徴として、その解析処理の中途、表層解析とより深い解析との間に明確な処理の断絶があることが挙げられる。実際の文法記述では解析全体を統一的に記述してあるが、処理アルゴリズムとしては、表層解析部分だけの処理までをすべて完了してはじめて、それ以降の解析・移行、生成処理を行っている。これは、解析のアンビギュイティの問題に対し、局所的に詳細な解析を行うことによって（詳細な情報を処理することによって）アンビギュイティを中途で解消していくという通例試みられている戦略を放棄していることを意味する。こうすることによって表層の解析において中間ノードに付与する統語素性の異なりが最小になり、「パック」の効果が最大になり、長大な文に対しても現実的な時間・空間的資源によって対応できているものと考えている。

アンビギュイティの問題に対応するために、我々はウエイトと呼ぶ評価値を最大限用いている。ウエイトは、辞書記述や文法記述、各レベルでの処理において与えられる。この値は不自然さ、非平明度、難解性、緊張度、情報性を表す「解析コスト」として扱われているので、値が少ないものの方が優先して選択される、あるいは値の大きいものは代替の解析の不可能なときに初めて選択されるようなものであると見なしている。これらの問題については稿を改めて議論したい。

### 1．2 局所解析の導入

原文で、固有名詞に関わる表現や数と時などに関わる表現をはじめ、比較的定型的な表現であって、文全体に関わるような構文規則には必ずしも馴染まないようなものを我々は局所解析（Local Context Parsing）、あるいは局所翻訳（LOcal Context Translation、ＬＯＣＴ）という処理要素を設定して取り扱うことにしている。ＬＯＣＴの導入には次のような自然言語システム構築上の有効性があるものと考えられる：

・人間が英語を読む際にも、大文字、数字そのほ

か幾つかの「キーワード」（ＬＯＣＴキーワード）を手だてとして、一連の単語列パターンを構文的に独立したひとつの単語あるいは句として認識していると考えられる。この仮定には、少なくとも自然言語処理の為の客観的妥当性があるであろう。

・そのようなパターンの多くは、表層の構文規則の中で一般的にも扱える。しかし、そのような規則を一般の構文規則の中に無制限に取り入れると、別の無意味な解釈が優先的に出てしまったり、追加した規則が（不注意に）別の誤った解釈を生んでしまうという危険性がある。

・そのようなパターンのより深い解析処理、さらにはターゲット言語の生成処理までには、通常の構文に対する扱いとはかなり性質の異なる処理が要求される。このようなパターンの処理内容だけをとれば、一般の構文的構造に比べても単純な場合が多いが、一般の場合に比べて「アドホック」で特殊な扱いが要求される場合も少なくない。

・そうした局所的定型パターンは一般の規則以上に分野や対象、時代などによって変遷するものである。

ここで述べた要因は、現在我々がＬＯＣＴの対象にしている範囲のものに止まるものではない。各種の慣用表現、ことわざ、句読法、字体や字下げ・列挙などの書法、挿入や引用、（拡張した意味での）係り結び等々、従来の言語論的な自然言語へのアプローチから外れる要因も少なくない。そのいくつかはＳｔａｒのアルゴリズムに無理を押しても吸収させているが、ＬＯＣＴの枠組みを更に拡大、それらの要因の多くをここに取り込むのもひとつの解決策であると考えられる。

現在のＬＯＣＴでは、英文中の一定のまとまりを単独の「単語」として認定、それに対する辞書記述を創出することをその任務としている。我々はその２つのシステム実現形態を持っている。その第１の形態では、ＬＯＣＴは辞書照合機能の内部に組み込まれる。我々の辞書照合機能には、複数個、複数種類の辞書を参照するほか、辞書未登録語に対する推定と辞書項目作成やハイフン語の分解処理なども含まれているが、そこにＬＯＣＴによる単語認定までを含める。辞書照合機能は、文字列に対する単語認定の全ての要求に答える。

第２の形態としては、ＬＯＣＴ処理を単語認定処理と表層の解析の間におき、単語認定によって得られたＷＦＳ構造、その中には局所解析用辞書データも含まれるが、その構造から出発してＬＯＣＴのパージンングを行うものである。

前者は主に字句処理をベースにしたパージングが中心になるが、単語認定という観点からは統合された扱いが出来るという利点を持つ。それに対し後者の場合には、辞書照合などにより得られた情報までが利用でき、更にＬＯＣＴ処理を表層の解析の第１段階として捉え、「単語」の枠を超えたところまで拡張することが出来るという利点を持つ。我々は後者の手法から出発し、前者の手法に移行したが、現在は再び前者の方式に復帰しようとしている。

Ｓｔａｒの解析の前段までの処理はすべて、同じＷＦＳ構造を作成・拡大・操作するものと考えられる。従って将来への展望としては、これらを１つの完結したモジュールあるいは「層」として措定することが概ね可能であると考える。そうであれば、テキスト処理、辞書照合・未知語処理やＬＯＣＴ処理、表層の構文解析などを統一的なＷＦＳパージングの枠組みの中で有機的に結合する可能性を検討することが出来る。

## ２．　局所解析の対象

局所解析において扱うべきパターンの種類は、分野やテキストの種類、特性によって様々に変化する。とくに各種ニュース記事などを翻訳対象に選んだ場合、人名、組織名などの固有名詞、日付や数の表現などに対する特別な扱い方が必要である（［２］，［３］）。ここでは固有名詞表現と数や時の表現の両者につき、そこでどの様な問題があり、どの様にそれを捉え処理すべきか、現在までの検討の概略を報告したい。

### ２．１　固有名詞表現

固有名詞表現を特別扱いしなければならない理由には次のようなことがある：

・それを１まとまりの名詞として扱わなければならない。それが、構文的にいくつかの独立した要素に分離されると、全く意味が捉えられなくなる危険性がある。特に、前置詞、接続詞などの一般機能語を含むとき、甚大な影響を引き起こす。

・固有名詞表現のすべてを予め登録しておくことは不可能であるが、その様な単語に可能な限り適切な訳語や意味的素性を与えたい。

・一般に固有名詞表現を多く含む文は長大なものであり、その解析上の負担もさることながら、莫大な解釈の発生の中で、人間にとっては受け入れ難い解釈が先行してしまう可能性が高い。

そこで、まず大文字で始まる語の並び（以下「大文字語連」）を語単位として捉えることから始めよう。語単位とは、その中に色々の構造を内包しているにしても、外に対しては１つの閉じた単語として扱っても構わない様な単位であることを意味している。

#### （１）　大文字語連

大文字語連の認定自体は、機械的に字面だけからでも行なうことができる。しかしそうした大文字語連を１つの完結した固有名詞単位とするのには次のような問題がある：

・大文字語連をとるときに、最長のものを採用するのか、あらゆるまとまり方を採用するのか。その時に「ウェイト」をどの様にコントロールするのか。実際に、大文字語だけの並びであっても、それを幾つかの語単語に分割しておかなければならないような場合も多い。
・また、連なりの長さとその尤らしさの関係はどのようなものか。これは、それが人名か、組織名か、地名かなどの要因によっても変化する。
・大文字語連がシステムが認識できる語、例えば辞書既登録の語などを含むときにそれをどう扱うのか。一般的には、そのような語の前後で大文字語連を切断してしまうことはできない。またそのような語も含めて大文字語連をまとめた場合、その全体の構造を正確に把握することなしに、その認識された語にだけ訳語を埋め込んでよい保証はない（例えば、苗字か名前の一方だけがカタカナで、残りが英字のままであるのも異様である）。
・文頭における大文字化をどの様に扱うのか。とりわけそれが辞書に小文字で登録されているときが問題であり、それを含めて１単位として扱う解釈は相対的に使いにくくしておく必要がある。
・類似の問題として、見出し、表題などでは、文頭に限らず大文字化が起きる。そのような環境では大文字連の尤らしさを減らしておく必要がある。
・引用符（人名の中の愛称など）やハイフンなど特殊記号への考慮も必要である。

大文字語連の中を観察してみると、そこにはその並び全体の性質とその処理方法自体を指示するような「キーワード」が含まれている。例えば、外電ニュース記事１カ月分の中で、大文字語連のパターン 30766種を抽出、その最後の単語毎に、それが持つパターン数を集計した。その上位４０番目までを表１に示した。なお、ここでは、語末のピリオドは無視してある。

| | | | |
|---|---|---|---|
| 249 | Party | 93 | Front |
| 159 | Center | 89 | Monday |
| 145 | Corp | 87 | Court |
| 143 | Committee | 84 | Tuesday |
| 137 | University | 83 | Jr |
| 136 | Minister | 81 | Group |
| 135 | Co | 80 | Friday |
| 135 | Association | 76 | Commission |
| 126 | Inc | 72 | City |
| 117 | Council | 71 | Sunday |
| 115 | Cup | 71 | Institute |
| 105 | Open | 67 | U.S |
| 103 | Wednesday | 66 | Federation |
| 103 | Hospital | 65 | League |
| 102 | Bank | 63 | Army |
| 98 | Agency | 62 | Service |
| 96 | Union | 60 | Smith |
| 96 | Department | 58 | Congress |
| 95 | Club | 56 | Stadium |
| 93 | Thusday | 56 | Office |

表１　多くのパターンで用いられる大文字連の最終語

ここでみても、「ＬＯＣＴキーワード」がかなり大文字語連全体の特性を与える可能性があることを示唆している。しかし、このような「キーワード」とおぼしき語は少なくとも数百個は準備する必要がある。

（２）　大文字語以外を含むまとまり

大文字語連だけではまとまりとしては閉じていないことも多い。その殆どは少数の機能語類によるものであるが、そのいくつかの要因を挙げてみよう：

・並列の "and" は、字面上の大文字語連を完結しないものにする可能性がある。取り込むならば、大局的な並列とのアンビギュイティを考慮しなければならない。"&" も "and" に類似しているが、通常これは大局的には用いられないものと考える。
・"of"、"for"、"on" などの前置詞も固有名詞表現のまとまりの構成要素になる。しかし、殆どの場合は局所解析でのまとめ上げの中でもより上位のレベルで扱い得るものと考える。
・人名などで、"von"、"van"、"al"、"de"、"di"、"d'..." などの前置詞が固有名詞単位に挿入あるいは先行することがある。"William of Occam" の "of" のように、英語の機能語が用いられているときの扱いには議論があるが、それを除けば「裸の固有名詞」の構成要素としてよいであろう。
・冠詞の "the" をまとまりの中に組み込むべき場合も多い。ただし、問題の大文字語連が主名詞に先行して並ぶような場合、冠詞がどちらを修飾しているかは明らかではない。
・コンマは、構文的な、局所解析の範囲を超える用法の他に、固有名詞表現などとしての同格（肩書、所属、所在地など）としての用法も多く、それらは局所解析の中での大局的な要因であると云うことも出来る。一方、会社名での "…, Ld." と

いった下位レベルでまとまるべきものについては、現在は ", Ld." 等をキーワードとして扱っている。
・そのほかにも、社名での "… & sons" といったいくつかの定型パターンが存在する。

### （３） キーワードなどによる固有名詞の構成

固有名詞表現全体の形式的記述は、大文字語連や固有名詞辞書データなどと局所解析キーワードを含む規則体系であるが、ここでその全体系を論ずることは出来ない。ここでは、人を指す固有名詞を例にとって述べることに止どめる。

非常に粗く描けば、我々は人名表現の構造を次のように捉えている：

＜人＞::=（＜役職＞）（＜称号＞）＜人名＞
（，＜役職｜所属＞，）

ここで＜人名＞とは、その全体が「人名」として辞書に登録済みのもののほか、大文字語連などで人名としての可能性のあるものを含む。人名の構成要素（姓、名、"von" の類、"Jr." の類など）がキーワードとして機能することもある。

＜称号＞は、"Mr."、"Dr." のようなものであるが、"…, Ph. D." のように後置するものも実際には扱う。

＜役職＞は "Chairman"、"U.S. vice-president" などで代表されるものである。これ自体、国名や機関名などを含むまとまりであることも少なくない。これが長い単位である時は、それを並置的な扱いとして、日本語訳もその順に訳出する方がよいこともある。

＜役職｜所属＞と記したのは、その前の人物を補足説明する種々の表現である。ここには組織・機関名などを含むきわめて多様な表現が用いられる。こうした同格表現が、人名の並列などに際して用いられることも少なくない。これは構文解析上きわめて大量のアンビギュイティを発生させ、解析の負担を過大なものにしかねない。その意味では、できうる限りＬＯＣＴのレベルでこのまとまりまでを解決しておきたい。

## ２．２ 時と数の表現

４万数千文、百万語以上の外電ニュース文から時と数のパターンを抽出、それを１１０程度に分類してその出現頻度上位４８位までについて、頻度とパターンの例とを併記したリストを表２に示す。なお、ここでは、経済記事の多くは対象記事から省かれている。

この統計において、パターンの選定には多分に恣惟的な要因も入り、他方では、出現のカウントに機械的な「ゴミ」も排除しきれていないとは思われるが、時と数とに関わる表現形態の概略を知ることはできよう。

| | | |
|---|---|---|
| 5674 | : | Sunday |
| 4133 | : | 1989 |
| 2377 | : | about 100 |
| 1329 | : | on Saturday |
| 1230 | : | On Feb. 2 |
| 869 | : | 30 |
| 669 | : | more than 380,000 |
| 663 | : | 27 years |
| 631 | : | At least six banks |
| 628 | : | 30-year |
| 568 | : | Jan. 31 |
| 479 | : | 18-year-old |
| 303 | : | in April 1991 |
| 288 | : | May 30, 1987. |
| 280 | : | 6.0s |
| 231 | : | 300 miles (480 kilometers) |
| 219 | : | since 1967 |
| 193 | : | in the eighth |
| 190 | : | the 138-year-old |
| 172 | : | March 18 |
| 159 | : | Since Jan. 20, |
| 156 | : | eight months ago |
| 146 | : | some 30 universities |
| 138 | : | on July 19, 1979 |
| 133 | : | the Feb. 4 |
| 131 | : | in his 40s |
| 124 | : | last Friday |
| 104 | : | a 1986 |
| 104 | : | Feb. 26-28 |
| 103 | : | for 6 p.m. |
| 102 | : | in the 1980s |
| 99 | : | on Sept. 22, 1927 |
| 93 | : | third in |
| 80 | : | at 1:59 |
| 50 | : | in the Feb. 25 general elections |
| 46 | : | about 300,000 dollars |
| 42 | : | 12-hour |
| 40 | : | until 1988 |
| 33 | : | 2-month-old |
| 30 | : | last Aug. 18 |
| 30 | : | nearly 80 millIion |
| 28 | : | in 1986 and 1987 |
| 28 | : | forty-five years after |
| 27 | : | Last Sept. 25 |
| 27 | : | after weeks of |
| 26 | : | on Feb. 24-25 |
| 26 | : | in 1961-62 |
| 24 | : | 88 years old |

表２ 時と数の表現パターン

### （１） 時の表現

時の表現は、単独で、あるいは前置詞句などの形で、さらには副詞節などの形で構文的にかなり大局的な機能を果たす場合が多い。この点においては場所の表現などとも共通する面があるが、構文的には時の表現に、より幅広い用法がみられる。特にニュース文では時の表現が非常に多様な形で用いられている。例えば次のような表現はニュース記事のほとんどにおいて見られる：

"He said (on) last Wednesday (that) ····"
"···· the fighting broke out Jan. 30"。

こうした、時の表現は、ニュースなどでは文中のほとんどあらゆる場所に出現し、構文規則としてもこれを特別に扱う必要が認められる。これは主に、時の表現の構文的役割に関わる問題である。

次のようなフレーズを示されただけで、殆どの場合、人はそれを時の表現であると解釈するであろう:

"in 1989"、 "in 1961-62"、"2001"。

こうした時の表現をそれと認定することが、大局的な構文解析に先立って行なうべき局所解析の主要な任務の一つであると我々は考えている。しかし、時の表現の認定は、今の例にも見られるように、決定論的に行なう訳にはいかない。また、その認定メカニズムも個別に複雑な規則になる場合が多い。

・"Jan. 30" はおそらく「１月３０日」であるが、"Jan. 1960" は「１９６０年１月」であろう。
・"1970s" は「１９７０年代」であるが、"30s" は「３０代」、「３０度台」、「３０年代」など３０～３９を示すどの訳語になるかは分からない。
・"3:30" は通常は時点の「３時３０分」だが、ボクシングなどでは「３分３０秒」の時点または時間の意味になることもある。
・"five to eight" を「８時５分前」とすることは余りに危険である。

時の表現は、"last"、"after"、"later" など、いくつかの形容詞、副詞によって前置／後置修飾され、まとまった一つの単位として機能する。このときにどの様にその構文的役割が変質するのかという問題を含め、微妙な問題を有している。

前置詞などとの関わりも微妙であり、より構文よりの分析が必要であるが、時点と期間、時の広がりといった演繹的な概念設定だけで対応できるとは限らない。多様な用例の中から、前置詞と時の表現との結び付きの例を１つだけ挙げておこう:

"(calls by ～) for a cease-fire for 6 p.m. and another for 11 p.m."

（２） 数に関わる表現

数に関わる表現は、（１）に述べた「時に関わる表現」の一部をも含んで、きわめて多様な広がりを見せる。経済分野をはじめとして、数字や数字を含んだ色々の定型的表現の羅列は人間にとって必ずしも読み取りにくい（聞き取りにくい）ものではないが、その約束ごととしての「型」を知ることなく機械的に処理しようとすると、過大な負担を機械処理に与え、しかもその結果として正しい解釈を得ることはきわめて困難である。そこで、局所解析でこうした部分を扱えるようになることは、自然言語テキストの機械処理にとって非常に大きな救けとなるであろう。ここではその全貌を示すまでには到底及ばないが、数に関わる表現の扱いに関連するいくつかのトピックを羅列していくことにする。

数そのものの認定（と計算）には、次のような留意点はあるものの、対処可能な範囲である:

・"and" を含む数の表現についてはアンビギュイティを考慮しなければならない。また、"Ten and three is 13." を、「１３は１３である」としてしまっては意味がない。
・"a"、"one"、"second" などについても、一般語とのアンビギュイティがある。また、"billion" が「１０億」（米）なのか「１兆」（欧州）なのかといった問題もある。
・"half"、"quater" などについても多少特別の考慮を要する。また、序数などの考慮も必要である。
・"$100 million" のような表現も計算する必要がある。
・"1.23 million" は「1,230,000」より「123 万」の方が読み易いであろうが、どの様な基準でそれを選ぶか。数字を全角・半角いづれで表示するか。
・ハイフンを含む数字表現をも処理する必要があるが、数字の修飾先までがハイフンで結ばれていることもある。

数及びそれを含む表現に対しては色々の要素が修飾する。直接、数の直前に現れるもの、"about"、"more than"、"up to"、"at least"、"nearly" などでも比較的扱い易いものであろう。しかし、それが数を直接修飾するとみるか、主名詞やフレーズ全体を修飾するとみるかは微妙であり、それによっては原文の意図を反映することが出来なくなることもある。

数の表現での後置修飾は、数とそれとの間に色々の要素が入り得るという意味でより扱いにくい。（１）の最後で述べた "twenty years later" のような時の表現もこの特別な場合である。更に、"per million gallons"、"··· a day" などは、その前が純粋な数量単位であれば単位の合成処理として扱えるが、これらの句は一般的には形容詞句として、あるいは副詞句として機能すると考えざるを得ない。しかし、これらが数の表現と関連を持って現れていることを無視して解析することはできない。局所解析でこのようなパターンをも認定すると同時に、適切なまとめ上げと情報の付

与とを行って、構文の解析に備えなければならない。

"12 cups of (coffee)" に対して、「１２杯の」といった形容詞を作り出すこともＬＯＣＴでの数に関わる表現の処理の１例である。

数に関わる表現には、"3 meters by 5 meters" のようなものにはじまる様々の定型的表現があり、スポーツ記事、経済記事など、対象によっても多様な形式を持つ。数に関わる表現はまた、時の表現におけるように、副詞的な動作などをすることも少なくない。

## ３．　局所解析の実現

ここでは、新しいＳｔａｒ翻訳エンジンにおける局所解析処理がどの様に実現されているかをアルゴリズム的な観点から述べる。そのために、ＷＦＳパーザの一般的な概念を準備する必要があるので、厳密な定義などは避けて直感的に把握されるよう、かいつまんで解説する。つぎにそれを局所解析のためにどの様に限定、具体化したかを述べる。そして最後に局所解析のための辞書についてふれることにする。

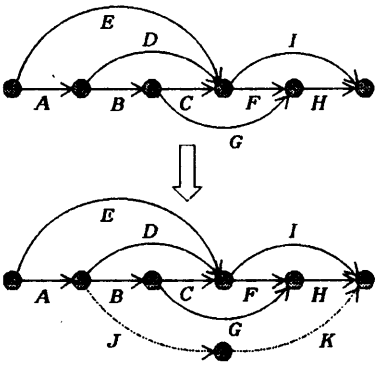
### ３．１　ＷＦＳパーザ

翻訳エンジンＳｔａｒの表層解析がＷＦＳ構造を基礎にしていること、そして局所解析が多様な解析操作を要求していることの２つの要因を考慮して、我々はＳｔａｒの局所解析にもＷＦＳパーザを採用することにした。ＷＦＳパージングによる表層構文解析の詳細については［１］に解説してある。ここでは（ＷＦＳ）パージング（表層の解析）のより一般的なモデルから概説することにする。図１に、我々がＷＦＳ構造と呼ぶものの簡単な例示をし、以下の説明でそれを用いる。

ＷＦＳ構造は、「語位置」と呼ぶことが出来るノードと、ＷＦＳと呼ぶ（有向）アークとからなるネットワーク構造で表現する。図１における「Ａ」から「Ｋ」までのＷＦＳは、それぞれ単語や統語成分に対応していると考えてよい。

ＷＦＳパージングは、図に示すように、それぞれのＷＦＳ構造に対しＷＦＳ部分構造を追加していく操作の繰り返しである。図１は、部分列「Ｂ・Ｇ・Ｈ」に書換え規則を適用して部分列「Ｊ・Ｋ」を追加するという簡単な例である。ここで「ｓｙｍ（Ｂ）」は、Ｂの書換え規則用のシンボルを表わすものとする。ここで分かるように、ＷＦＳパージングはボトム・アップ型に処理を進める。

次の問題は操作の順序である。我々のインプリメンテーションでは、あるＷＦＳに着目し、それより右にあるＷＦＳはすべて作成済みであるとの仮定の下に、そのＷＦＳの始点ノードをルートとして右に延びるような部分構造に限って作成・追加するものとしている。




規則：sym(Ｊ)･sym(Ｋ) → sym(Ｂ)･sym(Ｇ)･sym(Ｈ)

図１　ＷＦＳ構造の拡大

そして、右から左へと着目ＷＦＳを移して同様の操作を続ける。勿論、向きをすべて逆転すれば左から右への解析を行なうことになる。図１は、実はＷＦＳ名の与え方からも分かるように、左から右へ解析が進むことを想定している。図の状態は、ＨというＷＦＳが着目されているときの様子であるということが出来る。

ＷＦＳパージングで用いる情報（単なる書換え規則なら、終端／非終端記号）が同じであるＷＦＳは同一視することが出来る。ただし、作成過程（統語木など）を後で利用するためには、ＡＮＤ／ＯＲグラフのような表現を用いてそれを記録しておく必要がある。逆に、ＷＦＳの追加の度に元のＷＦＳから訳語などを作成し、それだけが目的であれば履歴を残す必要はない。そのかわり、ＷＦＳ同一化の際にはそうした項目まで含めて区別する必要がある。ただし、ウェイトのような項目は、その最小値をＷＦＳの値とすればよい。

### ３．２　局所解析用ＷＦＳパーザ

Ｓｔａｒの表層解析用パーザは、効率上の問題から、規則の表現力を必要十分な範囲に押し込めて最適化した専用プログラムであるということが出来る。これに対し局所解析の場合、ＷＦＳ生成の影響が概ね局所的な範囲だけにとどまるものと考えられ、パターン判定やＷＦＳ作成に様々な定型化できない処理が残っていることも考慮すると、効率よりも柔軟な表現能力の方が重要である。現在の我々の局所解析では、Ｃ言語の手続き型記述を含んだ規則記述を、プリプロセッサによってＣ言語ソースに変換、利用するものとしている。

それぞれの規則記述は、着目ＷＦＳに対してその後続ＷＦＳ列を検査するものである。ＷＦＳパーザの実現形態の概要を知るために、動詞句の単純な規則を例にとって、規則記述方式の概念的な例示をする：

ルート・ノード：　（Ｃ言語の関数入口）
　自動詞なら、動詞句ＷＦＳを作り、
　　ノード２に進む.
　他動詞なら、ノード１に進む.
ノード１：
　名詞句なら、動詞句ＷＦＳを作り、
　　ノード２に進む.
ノード２：
　前置詞句なら、動詞句ＷＦＳを作り、
　　ノード２に進む.

ここでの規則は、ＷＦＳ列が与えられたとき、それが条件に適合するか否かを検査するものである。実際のＷＦＳ構造はネットワーク型の構造であるが、規則記述においては列が与えられると考えていてよい。

それぞれの規則のノード毎に、前のＷＦＳに後続するカレントなＷＦＳが対応する。規則記述に当たって、そのＷＦＳだけでなく、これまでに採択したＷＦＳ列、規則内で設定した変数、様々の環境の情報なども利用することが出来る。又、ウェイトのコントロールもここで行なう。

ある規則ノードでＷＦＳ部分構造の作成を行なうと、ルート・ノードにあたるＷＦＳの始点を始点とし、カレントなＷＦＳの終点を終点とするＷＦＳ部分構造をＷＦＳ構造に追加することになる。なお、この節での例示は、右から左への解析を想定した（図１とは逆）。

実際の局所解析の記述の多くは、次の程度の表現力で十分である。これらを上記の規則プログラムの形に書き直すことは容易である。

```
rankof ::= PRESIDENT(x) OF organization(x)
       ::= CHAIRMAN(x) OF organization(x)
    ; [$2 の $1]
```

ここでは、素性変数ｘを利用することによって、「…の大統領」、「…の社長」、「…の会長」、「…の議長」などの振り分けを行なおうとしている。

局所解析の最終段階では、完成したＷＦＳ構造の中から必要なＷＦＳだけを選び取り、そこから新しい辞書項目を作成、それまでの単語認定の結果（解析用ＷＦＳ構造）にそれらを追加する。現在、局所解析の内部でのＷＦＳ構造はそれ以降の処理には用いられない。

### ３．３　局所解析用辞書データ

局所解析の「終端記号」としての「単語」は、数字や大文字に始まる未知語など直接字句処理により得られるもの以外は、局所解析用辞書データとして予め準備する。局所解析用辞書データはＳｔａｒのオプション辞書形式で作成するので、専門語などの場合と同様に、局所解析用辞書データを含むオプション辞書を任意個適切なものを選んで翻訳処理に渡すことが出来る。

局所解析用辞書データのうち第１のタイプのものは、ＬＯＣＴキーワードと呼ぶべきもので、局所解析におけるパターンの認定と訳語の作成に必要な情報を与える。そのような用語は非常に広範囲にわたっており、"Mr."、"Co." のようなものから、"Prime Minister"、"Club" など、さらには "wide"、"early" などの一般語までを含む。ＬＯＣＴキーワードは原則として単独で翻訳処理に使われることはない。

第２のタイプは人物名、組織名、地名などに代表される固有名詞類そのものである。これは他の局所解析用単語と結びついてひとつのまとまりを作るほか、単独でも翻訳処理の対象になる。従って、このような語は原則としてＳｔａｒのシステム辞書には登録しない。

何れのタイプの用語の範囲も、翻訳対象によって大幅に変わってくる。特に後者では、アップ・ツー・デートに更新すべき要因が大きい。この為の管理・運用方法はさらに研究すべき課題であると考えている。

## ４．　おわりに

英日翻訳システムを対象に、局所解析という処理単位を設定、その概観を論じてきた。しかし、局所解析の対象パターンに関する分析・実験の報告など、その具体的内容にまで触れることが出来なかった多くの事項がある。これらについては稿を改めて報告したい。

これまで自然言語処理においては、言語論的なモデルに基づくシステム構築が議論の中心になってきたようにも思われる。しかし、ここで提示した問題を含む「泥臭い」問題を統一的に解決していくための議論は自然言語処理の実用化のための核心的課題の１つであると我々は考える。拙論がこうした意味での何がしかの貢献をもたらすことが出来れば望外の喜びである。

最後に、この研究において加藤直人氏をはじめとするＮＨＫ放送技術研究所の方々に多大な協力を頂いたことに対し謝意を表したい。

［参考文献］

［１］中瀬：「英日機械翻訳システムにおける解析手法について」，情報処理学会自然言語処理研究会資料６９－７，１９８８.
［２］相沢ほか：「衛星放送ワールドニュースの英日機械翻訳」，情報処理学会第４０回全国大会２Ｆ－１，１９９０
［３］加藤ほか：「英日機械翻訳における固有名詞処理」，情報処理学会第４０回全国大会２Ｆ－２，１９９０