知識工学と人工知能　28－2
（1982. 12. 8）

マイクロコンピュータ上での補完的英文生成システムの実現

松永義文・小川均・田中幸吉（大阪大学）

1. はじめに

英文を生成するには、どのような"意味"の英文にするのかという情報が必要であるが、現在の我が国における計算機環境というものを考慮した場合、この"意味"を与える媒体としては『英語』が最適と思われる。入力の容易さや英語の普及度などがその理由である。英語の入力に対し、英語の出力という一見矛盾した考え方のようであるが、ここで意図する入力英語というのは、生成して欲しい英文のための最小の基盤的情報である。システム開発にあたっての背後的思想とは、おおよそ次のようなものである。

「あなたの知っている英単語で、あなたの知っている文法や表現で、できるだけ自然文に近いとあなたが思う範囲内で、あなたの作成したい英文の意味を英単語の並びとして表現して下さい。あなたの気持を名詞の羅列でもいいですから入力して下さい。そうすれば、その意志を酌みとってあなたの入力した英単語を使って、より豊富な情報を有した完全で文法的に誤りのない英文を作ってあげましょう。」

この思想の完全なる実現は困難であるが、本システムはその実験的試みとなっている。

本論では、筆者らが開発を進めているこの補完的英文生成システムについて概説を行なう。2.ではシステムの特徴について、3.ではシステム各部の説明を、4.では動作例を紹介する。

2. システムの特徴

結果的に本システムは、英文作成を支援するもので、使用者の期待する英文へのベースとなる英文を構成して提供するサポータである。システムの特徴を整理すると以下のようになる。

1) 入力の不完全さを想定している

使用者としては、より少ないメッセージ(入力)で済めば、手間が省けて重宝である。コミュニケーションにはフルセンテンスは必要でなく、断片であってもそれが自然な並びであれば十分である。これは、フルセンテンスや何らかの制限を課せられた文の入力を前提としてきた従来のシステムとは思想的に大いに性質を異にしている点である。

(b) テキストを扱う

従来の自然言語処理システムの多くが1個の文のみを対象にしてきたのに対し、本システムはテキストを処理する。また悪文や誤文が含まれていても処理は進行し、入力テキストに対応した出力テキスト(使用者の意図に沿っているかどうかは別として)がいかなる場合にも得られるという柔軟性も特徴の1つである

(c) マイクロコンピュータ上に実現されている

現在に至っても本格的言語処理は、依然巨大な計算機環境を必要とし、これは、実用化や一般普及への大きな課題となっている。マイクロコンピュータという小規模環境にインプリメントされた本システムは、システムの性質・機能などから評価されねばならないものの、新しい試みであると思われる。

(d) Basic English を採用している

小さな記憶空間という制限を克服する手段としてBasic English(約850語からなる英語体系で、一般の会話を想定した場合には、Basic Englishで十分な表現が可能であることが示されている)を起用している。これに、換言表と専門用言表さらには未定義語処理機構を付加することによって、ひじょうに柔軟なテキスト解析態勢が得られている。したがって登録語数850という数字は、見かけ上入力にはほとんど制限を与えない。

3. システムの構成

本システムの概略の全体像は、図1に示されている。入力テキスト中の単語の検索に始まり、英文テキストを合成出力するまでにかかわる言語処理の各領域は広範囲となるが、本論では、特に後述の補完部と生成部を中心に説明を行なう。

(a) プリプロセッサ

本システムでは、入力形式は比較的自由で、たとえば、

句 ; 文 . 文 ?
句

というような形式も許可されている。また、次の3種 文 . と 文 , . と 文 はすべて許容の形式である。すなわち、構造上明確な限りにおいて、いくつの文(又は句)から成立しているかを認識する文単位セパレーションと省略形(I'm, can't など)の処理を含めた単語単位セパレーションという2レベルのフォーマライズィングを行なっている。

フォーマライズドテキストの構成要素は次の3つである。

1) 英単語　2) 特殊記号　3) 数字

このうち、1)の英単語に関してのみ辞書引きを行ない この段階で同時に各種語尾処理を施している。

プリプロセストテキストとは、入力テキストから得られたフォーマライズドテキストとそれに伴う辞書情報を合わせたものである。(→図2参照)

なお、辞書に関する説明はここでは紙面の都合上割愛させていただく。

(b) アナライザ

アナライザの仕事は、大別すると次の4つである。

1) 熟語の認識
2) 名詞句の認識
3) 動詞句の認識
4) 肯定単文への変換

たとえば、

Has he been in England for 2 years in all ?

という文に対し、

- 熟語 in all
- 名詞句 2 years
- 肯定文への変換
- 動詞句 has been

という解析を行なった結果

He be in England for 2-years in-all
+(疑問文という情報, 現在完了形であるという情報)

が得られる。これがアナライズドセンテンスである。これを品詞列としてみると、既に文の各成分の論理的格関係の機械的抽出が可能な形になっているのがわかる。

アナライザは、このような文の論理的な構造を解析するに都合のよい形式へと単語の並びを統合してゆく機能を持っている。なおここで言及している論理的格というのは、一般に言う英語の5文型

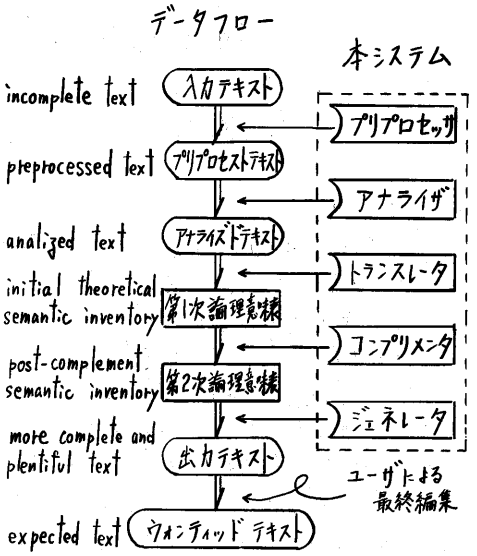


図1. システム構成と処理の流れ

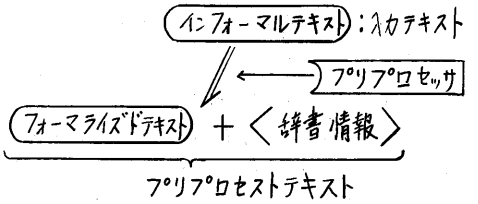


図2. プロプロセッシング

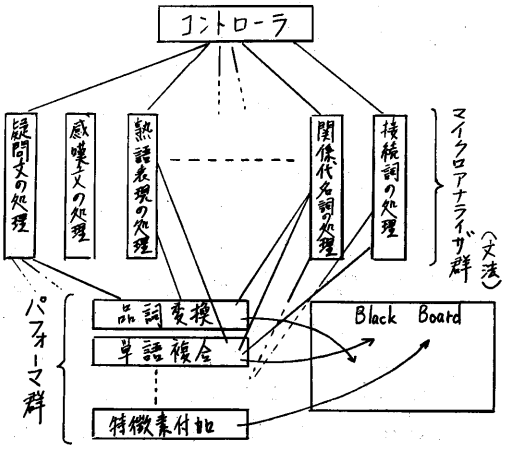


図3. アナライザの構造

に準じており、したがって本システムにおけるアナライザとは、すべての文を5文型のいずれかのパターンに帰着させるための一種のフォーマライザである。

アナライザのだいたいの構成は図3で示されている。テキストの中間結果は常にブラックボードに表示される。文法は小さな単位に分割されマイクロアナライザとして各々独立に記述されていて、おのおのの適用に関する制御はコントローラにまかされている。このような構成にしたのは、透明性と追加・変更の容易性のためのみならず、文解析における並列処理の必要性に答えるためである。すなわち、どのマイクロアナライザも複数回起動することによって見かけ上の並列処理を可能にできるわけである。

(c) トランスレータとセマンティックインベントリ

アナライズドテキスト中のどの文も肯定単文、または句であり、特に名詞句などは既に1つの名詞に変換されているので、論理的意味項目(5文型の思想に沿っている)の機械的抽出が可能である。この意味項目を表したのがセマンティックインベントリ(論理意味表…以後簡単のためSIと略記する)であり、図4のような構成になっている。

図5に示すように、アナライズドテキストからSIへの機械的変換と第1次補完(→次項(d)参照)を行なうのがトランスレータであり、その結果生成されたものを第1次SIと呼ぶ。

例として、過去完了形とわかっている次のようなアナライズドセンテンスのトランスレーションを考えてみる。

I	come	with	the-idea-of-sleeping	at	his-hous	that-night.
N	V	P	N	P	N	N

動詞があるから、その位置から左へ見て最初の名詞が主語である。PNは前置詞句であり文型要素とはなりえない。そこでPNはPの特徴素に従って所定の付属項欄に入れられた後削除され、NVNが残る。ここでVの指定する文型はこの場合comeに依存するのでSVという第1文型であり、最後のNは日(副詞)とみなされ、nightの特徴素から時の欄に入れられる。この結果第1次SIとして次に示すものが得られる。

主語S	----- I	場所PLA	----- his-hous
動詞V	----- come	時TIM	----- that-night
時制Te	----- 過去完了	手段WAY	--- the-idea-of-sleeping

トランスレーションは動詞が存在する場合は比較的危なげなく行なわれる。(動詞がないときの処

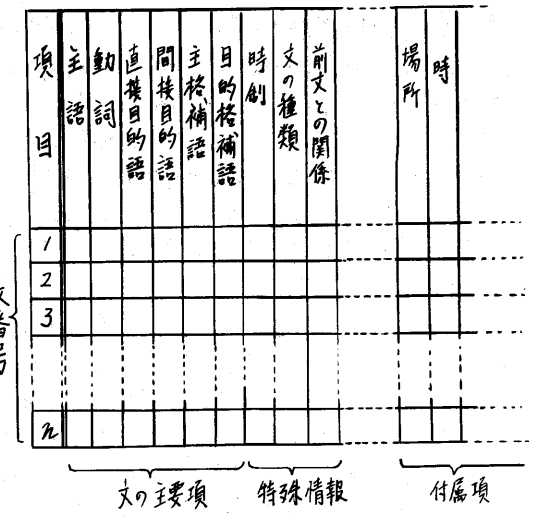


図4. セマンティックインベントリ

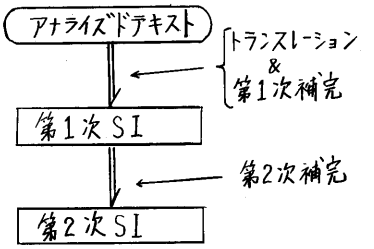


図5. 補完とインベントリの関係

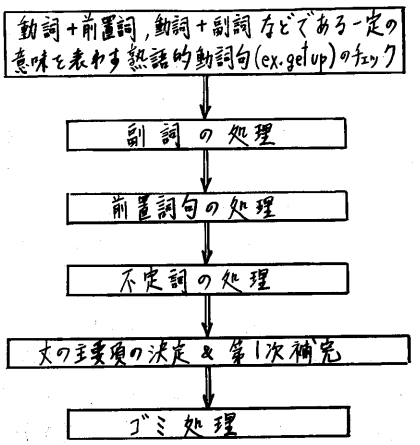


図6. トランスレータにおける処理手順

理手順の説明は省略させていただく。なお本例の場合、場所PLAのhis-houseはcomeという動詞とは関係がなくsleepingの行なわれた場所とすべきところであるが、このトランスレーションの段階では、場所の欄に入れられたというにすぎず、項間の係り受け等に関するチェックは、後述のジェネレータによって行なわれる。)
図6は、トランスレータの処理手順を示している。

(d) コンプリメンタ

トランスレータに出力された第1次論理意味表の欠陥部を補足し、より豊富な情報を推論によって付加することにより、より完全(あるいは冗長)な第2次論理意味表を生成するのがコンプリメンタである。

図5にも示すとうり、補完は、トランスレータの行なう第1次補完とコンプリメンタの行なう第2次補完とにシステム上で大別される。

第1次補完とは、

動詞が存在するが、その動詞の指定する目的語や補語が存在しなかった場合

直接目的語に対し	something
間接目的語に対し	someone
補語に対し	some-situation

を要求に応じて仮に補っておくというもので、これは、必ず補なわなければならない項を明確にするための補完である。

第2次補完とは、

第1次補完によって補なわれたものが具体的に何であるかを推論したり、その他の項を推論の可能範囲で補なうというもので、したがって、コンプリメンタの仕事とは、第2次補完そのものである。

図4に示すとうり、SIの項目は、主要項、特殊情報、付属項の3つに大別され、コンプリメンタは、このようなSI項を補足する一方、単語からの連想や項目間の関係から新しい文をも生成する。この機能は補足というより、付加的である。本システムにおける補完とはしたがって〔補足+付加〕の意味で使用している。

以上のことから、補完の体系を表現したものが、図7である。

図7は、形式的分類であるが、これらの補完を実現するための実際の方法論に基いて考案した補完法を模式的に図に表わしたものが図8である。補完法は、i) 文脈型補完法 ii) 連想型補完法 iii) 推論型補完法 iv) デフォルト型補完法 に大別される。各補完法とSI項との間の関係を表わしたのが図9

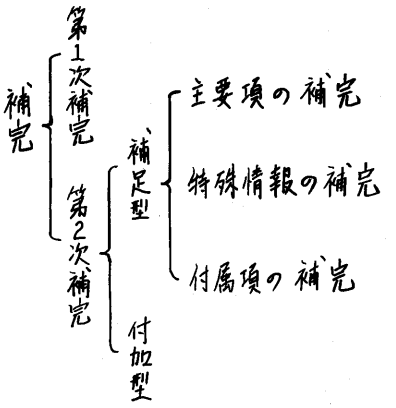


図7. 補完の体系

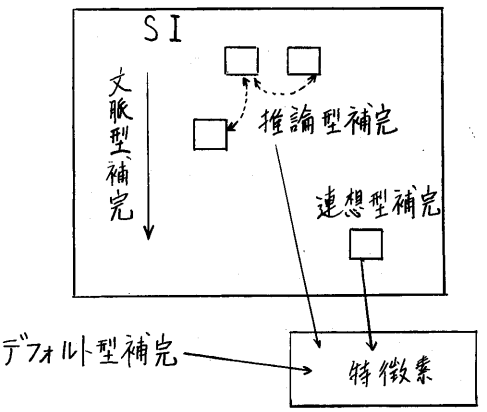


図8. 補完法

主要項
- S -- 文脈型, 連想型, デフォルト型
- V --- 文脈型, 連想型, 推論型, デフォルト型
- O --- 文脈型, 連想型, デフォルト型
- C -- デフォルト型

特殊情報 ---- 文脈型, デフォルト型

付属項 ------ 連想型, 推論型

図9. 補完法とSI項の補完関係

である。以下各補完法を簡単に説明する。

i) 文脈型補完法

テキストを処理対象としているため、前後の文の情報を参照することが可能である。主語が省略されていれば、それは前文のそれと同一である可能性が大きいし、また一連の文章であれば時制は連続(普通は未来に向かって)している可能性が強い。このような繰り返しを避ける目的で省かれた主要項や特殊情報を補完するものである。

ii) 連想型補完法

動詞と目的語の特徴素から、その単語に関連した情報を探索する方法である。たとえば、動詞句 get-up の特徴素から、時が朝で、場所がベッドないしはフトンという情報を引き出し、もしその文において時や場所の項が空であればそれらを補足しようというわけである。連想型には、したがって主要項と付属項のあらゆる種類を補完する可能性がある。

iii) 推論型補完法

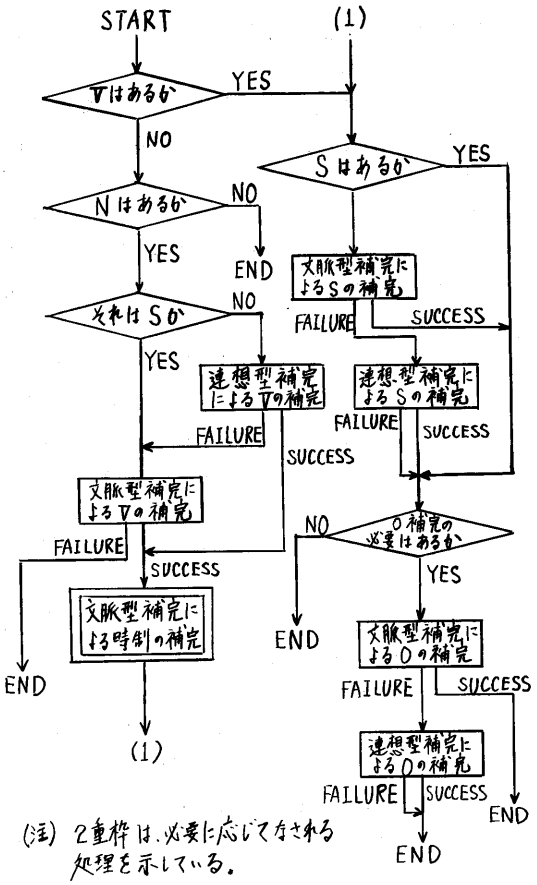
これは経験的に、項間あるいは文間の関係から推論によって補完を行なおうというもので、たとえば、動詞が get や have といった受領の意味を持つもので、目的語が present, girl-friend とかの受け取ればうれしいものであるとき、(主語) be happy という文を生成付加する、といったたぐいのいわゆる知識利用による推論であり、いくらでも多くの種類が考案できる可能性がある。

iv) デフォルト型補完法

i), ii), iii) のいずれの補完法によっても 動詞が決定されなかった場合適用するもので、特徴素の種類によって適宜処理される。

補完には、動詞の有無が大きな問題となる。また、目下のところ 補語 C を補なう手だてはなく、確立は困難である。主要項(Cを除く)と時制を補完するためのアルゴリズムは 図10に示されていて、このアルゴリズムを中心に設計されたコンプリメンタの構成を示したのが図11である。図11においての整合性のチェックというのは、たとえば 時の項に"yesterday"があるのに 時制の項に"現在"という記述があるとか、動詞が"give", 主語が"I"なのに 目的語が"me"であるといった 不整合を修正するものであり、これは第2次SIの理論的正統性を確保するために行なわれるものである。

デフォルト型補完法の導入により、コンプリメンタを通過すればいかなる入力であろうともフルセンテンスとなる。しかしながら目下のところ たとえば特殊記号だけとか、冠詞だけとかいうナンセンスな入力は無視



(注) 2重枠は、必要に応じてなされる処理を示している。

図10. 主要項と時制の補完アルゴリズム

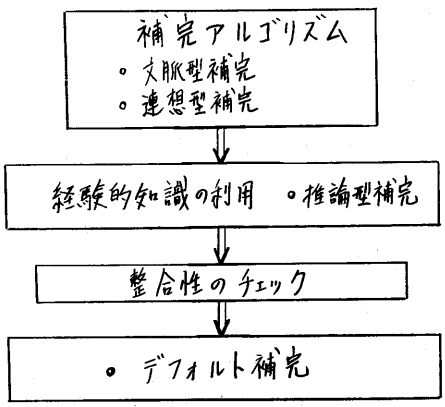


図11. コンプリメンタの構造

して処理の対象から除外している。

(e) ジェネレータ

ジェネレータはコンプリメンタの出力である第2次SIの情報をもとに、英文テキストを生成する。本システムでは、ジェネレータは次の2つの課題を負っている。

1) 文法エラーを含んだ文は絶対生成しない
2) 極力構文化し、よりエレガントなテキスト生成を目指す。

図12は、ジェネレータにおける処理の手順を示したものであるが、この図において特に<1><2><3><4>は、1)に、<9><10>は2)に関連している。

以下、図12の各部について概説する。

<1> 動詞のFormのチェック

指定された時制のFormに動詞を変換する。

<2> ST関係のチェック

俗に言われる"3単現のS"をチェックしたり、特に"have","be"動詞のForm等をチェックする。

<3> 冠詞のチェックおよび複数化

SIにおいて脱出であることがわかっているのに、冠詞として"a","an"といった不定冠詞が付いているときこれを定冠詞"the"に直す。また、普通名詞が単数形でしかも単独で存在している場合は、これを複数化する。

<4> 格Formのチェック

特に、人称代名詞についてのその論理格とフォームとの間の不整合を解消する。

<5> 付属項および付属項間チェック

たとえば、場所:school, 終点:school, 動詞:go というような状況は起こりうるが、このときは場所の項を抹消するのが妥当である。またたとえば、終点:school, 始点:home, 動詞:go というような場合には常識から考えて始点の項は不要である。このように同時に存在することが望ましくない付属項間での調整や著しく冗長性を増加させると考えられる付属項の削除を行なう。

<6> 特殊チェックI

間接目的語:someone, 直接目的語:a-camera, 方向:him なる状況は、たとえば"I gave a camera to him."という文を処理した結果得られることになるが、この場合は、間接目的語か方向のいずれかの項を削る必要がある。このようなおもに第1次補完の引き起こす不都合を処理する。

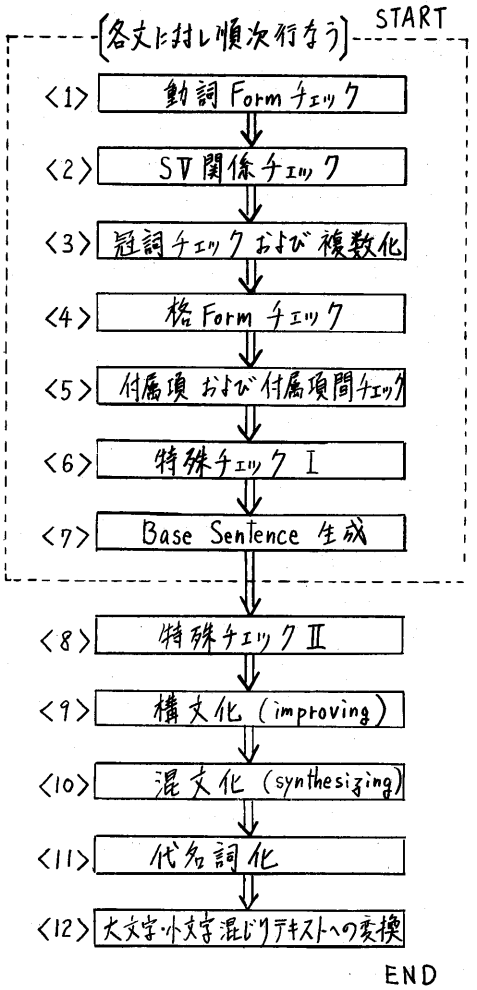


図12 ジェネレータにおける処理手順

<7> Base Sentence の生成

この段階で、疑問文,感嘆文,宣言文とそれぞれの指定されたFormに文を組み立てる。notなどもこの時挿入される。また付属項の位置は、存在する付属項の種類によって順序関係が決定される。

<8> 特殊チェックII

<8>以下は、テキスト全体を対象としている。ある文に、someone以外の間接目的語が存在し、その次の文の主語がこれと異なるしかもコンプリメンタによって補なわれたものである場合たとえば

(⟶次頁の動作例をとばして次々頁へ続く)

4. 動作例

入力テキスト

```
I GOT UP AT 7, AND SCHOOL AT 8
  MATHEMATICS AND ENGLISH
 AT HOME 4.                SUPPER
```

プリプロセストテキスト

```
I GOT UP AT 7 , AND SCHOOL AT 8
ﾋﾝｼﾍﾟﾂ = MVPPN,CNPN
(I M SIN PER SUB MO1 &ワタシ)(GOT V #04 (PAS &トッタ)(PPA &トル) SVO)
(UP P DIR &ノショウホウヘ)(AT P PLA TIM &ニ)(7 N NUM &7)(, , ,,,)
(AND C (AND &ソシテ)(PAR &ト))(SCHOOL N PLA &ｶﾞｯｺｳ)(AT P PLA TIM &ニ)(8 N NUM &8)
          ------------------------------------------
MATHEMATICS AND ENGLISH
ﾋﾝｼﾍﾟﾂ = ZCZ
(MATHEMATICS Z &MATHEMATICS)(AND C (AND &ソシテ)(PAR &ト))(ENGLISH Z &ENGLISH)
          ------------------------------------------
AT HOME 4 .
ﾋﾝｼﾍﾟﾂ = PZN.
(AT P PLA TIM &ニ)(HOME Z &HOME)(4 N NUM &4)(. . ...)
          ------------------------------------------
SUPPER
ﾋﾝｼﾍﾟﾂ = Z
(SUPPER Z &SUPPER)
```

アナライズドテキスト

```
1   I GET UP AT 7     : T10
2   SCHOOL AT 8     : , RPA
3   MATHEMATICS-AND-ENGLISH
4   AT HOME 4 .
5   LAST-MEAL-OF-THE-DAY
```

第1次SI

```
---------------  SEMANTIC NET  ---------------------
 1    S:I V:GET-UP TENS:T10 TIME(1): AT-7
 2    RELATION:RPA PLACE(1): SCHOOL TIME(1): AT-8
 3    O:MATHEMATICS-AND-ENGLISH
 4    PLACE(1): AT-HOME TIME(1): AT-4-O'CLOCK
 5    O:LAST-MEAL-OF-THE-DAY
-------------------------------------------------
```

第2次SI

```
------------------------  SEMANTIC NET ----------------------
 1   S:I V:GET-UP TENS:T10 PLACE(1):AT-HOME TIME(1): AT-7 STATING-POINT:
FROM-THE-BED
 2   S:I V:GO-TO TENS:T10 RELATION:RPA PLACE(1): SCHOOL TIME(1): AT-8
STATING-POINT:AT-HOME TERMINAL: SCHOOL WAY:BY-SOME-WAY
 3   S:I V:HAVE-EDUCATIONS-OF O:MATHEMATICS-AND-ENGLISH TENS:T10 PLACE(1):
 SCHOOL TIME(1):SOME-TIME COMMENT(1): ABS:DIFFICULT COMMENT(2):TIRED
 4   S:I V:GO-BACK-TO TENS:T10 PLACE(1): AT-HOME TIME(1): AT-4-O'CLOCK
STATING-POINT: SCHOOL TERMINAL: AT-HOME WAY:BY-SOME-WAY
 5   S:I V:HAVE O:LAST-MEAL-OF-THE-DAY TENS:T10 PLACE(1):AT-TABLE TIME(1):
SOME-TIME COMMENT(2):HAPPY
-------------------------------------------------------------
```

出力テキスト

```
I got up from the bed at 7 and went to school by some way at 8.
I had educations of mathematics and English that were difficult while I were at
school , so I got tired.
I went back to home from school by some way at 4 o'clock.
I had last meal of the day at table in the evenig , and I was happy.
```

(⟶前々頁よりの続き)

I met Tom at the station.
seemed very sleepy. (入力テキスト)

⇓

I met Tom at the station. (コンプリメント後)
I seemed very sleepy at the station.

⇓

I met Tom at the station.
He seemed very sleepy.

の例から明らかなように、第2文の主語は間接目的語に変えるべきである。これはコンプリメンタでの整合性チェック機能と類似しているがここでのチェックはそれより弱いもので、そうした方が正解の確率が大きいという場合である。これは、動詞の指定する主格の性質によってチェックされる場合であるが、その他テキストレベルでの経験的処理はここで施される。

〈9〉 構文化

いろいろの言い回しを形式化し、そのグレードに対して点数を与えたものを適用条件とともに登録しておき、各また対し適用条件を満たしかつ点数の大きい構文が発見されれば構文化する。特にこの処理の部分をimproverと呼んでいる。improverは、1文⟶1文, 2文⟶1文という2つのタイプの変換を可能にする。

〈10〉 混文化

極力単文化,複文化を行なう。この処理の依りどころとなる情報は、動詞の意味より与えられる場合もしくはもともと単文や複文であったというシンボルによる場合があるが、特に前者を実現するために、動詞(句)間の意味的対応付け表を用意している。

〈11〉 代名詞化

同じ名詞が連続して2つ発見された場合(ただし、2文間以内)後者の方の代名詞化を行なうというものである。

〈12〉 大文字・小文字混じりのテキストへの変換

本システムでは、簡便さという理由で入力はすべて大文字で行なう。したがって使用者へのサービスとしてこの処理を取り入れている。

ジェネレータは現在改良中で、ジェネレーションの一般的定式化(もちろん内部表現からのジェネレーションの場合である)に向けて考察を続けている。

5. おわりに

本論文では、マイクロコンピュータ上にインプリメントされた補完利用型の英文生成システムについて各部の概説を行なった。

システムとしての関係分野が広範であるため実際には現段階ではひじょうに多くの問題点が山積みしており今後改良に努めなければならないが、「自然言語処理」はそのどの部分を切り出しても依然巨大な問題であることは周知のことで、したがって当面は、研究の視点を定める意味で、次の2点を課題としてゆきたい

1) 補完技法の体系化と定式化
2) 英文生成の定式化への考察

6. 参考文献

1) L.W. Lockhart, 「Basic Picture Talks」 The Basic English Publishing Co.
2) 室 勝, 「新850で書く英語」 The Japan Times 社
3) 室 勝, 「基礎単語の使い方」 評論社
4) 室 他, 「別冊宝島14:道具としての英語,会話編」 JICC出版局
5) 松永他, 「補完手法を用いた英文生成について」 情報処理学会第25回全国大会, 1982