英文テキスト補完生成システムの補完能力評価†

唐沢　博††*　小川　均††　田村進一††

筆者らは，英文テキストの作成を支援する目的を持つ英文テキスト補完生成システムを既に作成した．このシステムは，作成したい英文テキストの要素であるような，単語・句・文といった断片的な情報を入力することによって，システムの推論の及ぶ範囲内で省略もしくは欠落している統語的・意味的情報を補い，利用者が意図していると予想される内容を持った英文テキストを生成する機能を持つ．この補完システムについて，次の2点の評価を行った．すなわち，(1)補完システムの補完操作の基盤である補完体系の特性を明らかにする，(2)補完システムの補完能力が人間の補完能力と比較してどの程度なのかを知る目安を得る．これらの評価の結果，(1)の評価によって，補完体系を構成する文脈型・連想型・推論型・デフォルト型の4種類の補完型のうち，内容的に質の良い出力テキストを生成する場合に連想型・推論型が寄与していることが判明した．また(2)の評価により，英語圏の人間の補完能力に対するシステムの補完能力を定量的に示すことができた．さらに，以上の評価の過程で副次的に明らかとなった評価手法の問題点や補完システムの諸特性に関して考察を加えた．

1. まえがき

筆者らは，人間が頭に浮かぶ単語・句・文などの集合からなる情報(以後，これを不完全テキストと呼ぶ)を入力すると，推論の及ぶ範囲内で必要な欠落情報を補って，文法的に完全な英文テキストを出力する機能を持つ英文テキスト補完生成システムを既に作成し報告した[1]．同システムの補完能力を評価する段階で，適切な一般的手法が存在しないことがわかり，補完能力の評価法を独自に考案した[2]．その評価法は，人間の頭の中にある意図とシステムの出力内容とを比較する必要から，心理実験的な手続きを含む方法となった．ただし，実験心理学においても本研究の目的に合ったような適切な手法は既にあるわけではなく，本研究のために新たに実験系を工夫した．補完能力の評価は，2種類の観点に立った評価法の適用により実施した．すなわち，(1)システムの補完操作の基盤である補完体系の特性を明らかにするための評価，および(2)システムの補完能力が人間の補完能力と比較してどの程度なのかを知る目安を得るための評価である．本論文では，英文テキスト補完生成システムの2種類の補完能力評価の手法と実施結果について述べる．

2. 英文テキスト補完生成システム

先ず英文テキスト補完生成システム（以後，補完システムと略称する）の全体的な構成と処理プロセス，および補完処理について概要を説明する．

2.1 システム構成

図1に示したように，補完システムは入力された不完全テキストに対し前処理を施してから，解析を行い深層構造へ変換する．この深層構造に対して補完処理が実行され，その結果を表層構造としての英文テキストに変換して出力する．補完システムの利用者は，この出力テキストにポスト・エディットを加え，目的とする最終テキストを得る．補完システムの目的は，利用者の英文テキスト作成の支援にある．**図2**に各プロセスにおける処理結果の例を示す．なおシステムは，Basic English (Ogden, C. K., 1930)[1] を対象とし，パーソナル・コンピュータ（富士通 FM 8）上にインプリメントされている．

2.2 補完処理

補完は，その段階において**第1次補完**と**第2次補完**とに大別される．前者は統語構造に基づく補完であり，統語構造上必要とされる項目に対し暫定的なマーキングを行う．後者は種々の知識を用いた意味レベルの補完である．第2次補完は**補足**と**付加**とからなる．これらは，(1)文の必須要素であるにもかかわらず欠落しているような情報を，統語的知識や一般的知識を用いて補足する手続きと，(2)存在すれば利用者にとって有益であろうと予測される情報の付加の手続きとである．これらを実現するために4種類の補完型が存在する．補完型は，1) 文脈型，2) 連想型，3) 推論型，4) デフォルト型があり，各補完型はプロダクション・システムに類似した機構のもとに規則形式の知

† Evaluation for the Ability of the English Text Complementing System by HIROSHI KARASAWA, HITOSHI OGAWA and SHIN-ICHCI TAMURA (Department of Information & Computer Sciences, Faculty of Engineering Science, Osaka University).

†† 大阪大学基礎工学部情報工学科
* 現在　京都教育大学教育学部教育実践研究指導センター

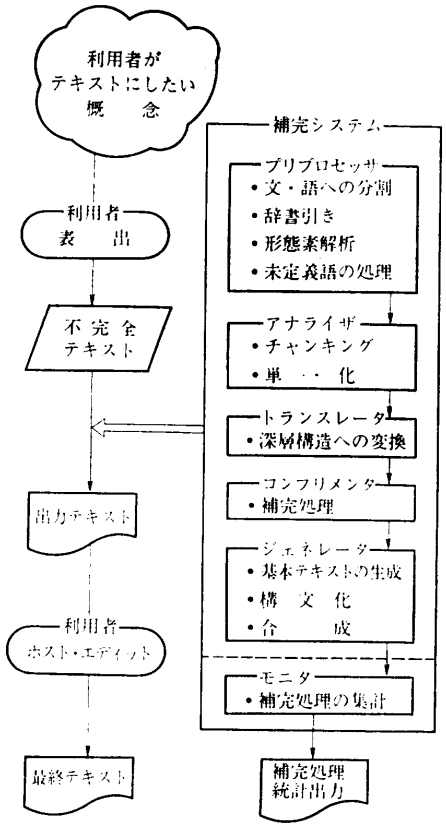


図 1 システム構成
Fig. 1 Construction of the system.

識によって構成されている．**表 1** に各補完型の特性を示す．

3. 評価 I ～各補完型の寄与率に関する評価

この評価は，補完体系の特性を明らかにするとともに，個々の利用者向けに tuning[3),4)] するためのデータを与える．そして，i）第2次補完を実行する4種類の補完型がどんな割合で寄与しているのかという点，および ii）質の良い出力テキストを生成するときどの補完型が効いているのかを知る目的をもつ[3)]．評価は，第1次補完および文脈型，連想型，推論型，デフォルト型を対象とした．

3.1 評価 I の実験系

図 3 は評価 I のための実験系である．補完システム

```
MADE A COOK OF EGG
HAD A BURN IN MY HAND
MOTHER GIVE HELP
```

（a） 入力テキスト

```
1. MAKE A COOK OF EGG
2. HAVE A BURN IN MY HAND
3. MOTHER GIVE HELP
```

（b） 前処理部の出力

```
1. MAKE A-COOK-OF-EGG :T10
2. HAVE A-BURN IN MY-HAND :T10
3. MOTHER GIVE HELP :T06
```

（c） 解析結果

```
1. V:MAKE  O:A-COOK-OF-EGG  C:in-some-situation  TENS:T10
2. V:HAVE  O:A-BURN  TENS:T10  WAY:IN-MY-HAND
3. S:MOTHER  V:GIVE  O:HELP  O:someone  TENS:T06
```

（d） 深層構造

```
1. S:I  V:MAKE  O:A-COOK-OF-EGG  C:IN-SOME-SITUATION  TENS:T10
2. S:I  V:HAVE  O:A-BURN  TENS:T10  WAY:IN-MY-HAND
   ADDITIONAL_ADJECTIVE:PAINFUL  COMMENT(0):CARELESS  COMMENT(2):SAD
3. S:MOTHER  V:GIVE  O:HELP  O:I  TENS:T10  COMMENT(0):GENTLE
```

（e） 補完された深層構造

```
I MADE A COOK OF EGG IN SOME SITUATION.
I HAD A PAINFUL BURN IN MY HAND BECAUSE I WAS CARELESS, AND I WAS SAD.
MOTHER GAVE ME HELP BECAUSE MOTHER WAS GENTLE.
```

（f） 出力テキスト

注） 1) T06 は現在時制を示すシステム・マーカ
2) T10 は過去時制を示すシステム・マーカ
3) 小文字の語句は第一次補完によるマーカ

図 2 処理例
Fig. 2 Example of processing.

表 1 各補完型の特性
Table 1 Properties of each type of complementing.

文　脈　型	冗長な繰り返しをさける目的で省略された情報を前後の文から復元する．
連　想　型	入力情報中の語をキーとして連想される情報を辞書中から取り出して省略を補う．もしくは冗長な情報として付加する．
推　論　型	一般的な知識（世界知識や常識的知識）を用いて，意味的に省略されている情報を復元する．
デフォルト型	上記3種類の補完型でも文要素が復元できなかった場合に，常識的な範囲で文要素を補う．

の動作原理を知らされていない試行者Aが表出した英語の断片情報を不完全テキストとしてシステムに入力する．システムは補完処理を行った後，テキストを出力する．もとの試行者Aはこの出力テキストを読み，自分の意図していた内容を表しているかどうかで，**質の良いテキスト群**と**質の良くないテキスト群**とに大別する．質の良いテキストと判断された出力テキストの

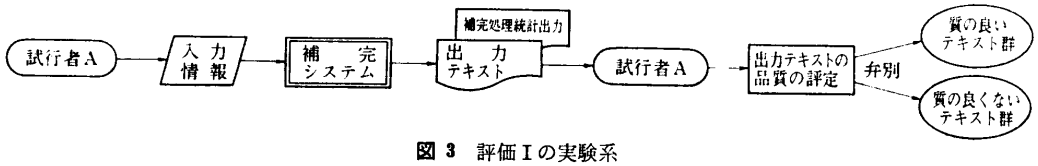


図 3 評価 I の実験系
Fig. 3 System for the evaluation-I.

```
SEE YOU AT KYOTO TOMORROW.
GIVE THE BOOK AS A PRESENT.
GET WHAT YOU HAVE IF POSSIBLE.
```

(a) 入力情報

```
I WILL SEE YOU AT KYOTO TOMORROW.
I WILL GIVE YOU THE BOOK WHILE I AM AT KYOTO.
I WILL GET WHAT YOU WILL HAVE FROM SOMEONE
   WHILE I AM AT KYOTO.
```

(b) 出力テキスト

図 4 質の良いテキストの例
Fig. 4 Example of an output text of superior quality.

```
WENT OUT MY HOME. AT STATION.
TOOK A TRAIN IN TIME.
TO TOKYO AND TOOK A MEAL WITH A GUIDE WHO WAS MY FRIEND.
```

(a) 入力情報

```
I WENT OUT TO MAKE SOMETHING.
I WENT TO MAKE SOMETHING BY SOME WAY TO STATION.
I TOOK A TRAIN AT STATION IN TIME.
I TOOK TO TOKYO AT STATION, AND I TOOK A MEAL WITH A GUIDE
   WHO WAS MY FRIEND WHILE I WAS AT STATION.
```

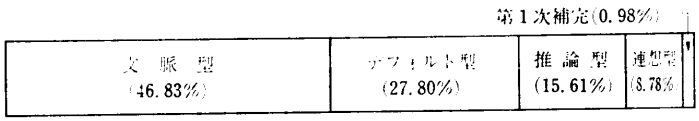
(b) 出力テキスト

図 5 質の良くないテキストの例
Fig. 5 Example of an output text of inferior quality.

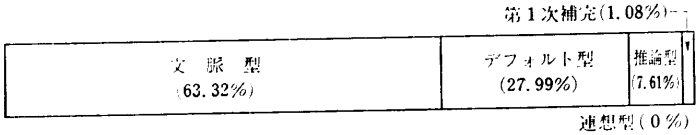
例を**図4**に，質の良くないテキストと判断された出力テキストの例を**図5**に示した．システムはまた，各補完型の起動回数と比率に関する統計情報も各出力テキストとともに出力する．

3.2 評価Iの結果

図6の(a)に質の良いテキスト群の結果を，(b)に質の良くないテキスト群の結果を示す．これらは，個個のテキストとともに出力される統計情報を各群ごとに集計したもので，各補完型の数値は，すべての補完型の平均起動回数に対する，個々の補完型の平均起動回数の百分率を示している．(a)と(b)とを比較することにより，推論型や連想型が多く寄与するほどテキストの質が向上すること，換言すれば人間の意を良く汲み取ることがわかる．なお試行者数は8名，1名当りの試行テキスト数は5題である．



(a) 質の良いテキスト群の場合



(b) 質の良くないテキスト群の場合

図 6 評価Iの結果
Fig. 6 Result of the evaluation-I.

表 2 辞書項目中の連想属性の例
Table 2 Examples of associative attributes in a dictionary.

語	連想属性
FROM	VER: COME
STUDY	CAU: INTERESTING
PAPER	PLA: ON-THE-DESK
LOVE	CON: BHAPPY ADJ: FANTASTIC
FEELING	VER: HAVE ADJ: SPECIAL
NEWS	CAU: CONSCIOUS CON: BSURPRISED
SWIM	VER: MAKE CAU: HOT

注) 1) VER: は連想される動詞
2) CAU: は連想される原因
3) PLA: は連想される場所
4) CON: は連想される結果
5) ADJ: は連想される形容修飾

この結果から，補完体系から見た省略の性質を考えてみると次のことが言える．すなわち，(1)文脈型補完が担っている文相互間の対比に基づく省略は最も多用される．(2)一方，一般的知識を用いて推論可能な物事や容易に連想されるような語・句・文の省略は，それらがうまく復元できたときの効果は大きい．(3)連想型補完は，その処理の中で辞書項目中の付属情報を用いるので，補完効率は補完ルールのみならず辞書項目の構成の仕方にも依存する．**表2**に辞書項目中の連想属性の例を示した．

4. 評価II～人間の補完能力との比較評価

補完システムの補完能力が，人間の補完能力と比較してどの程度なのかを知る目安を得るための評価である[3]．

4.1 試価IIの実験系

図7は，評価IIの実験系である．試行者Aが表出した英語の不完全テキストを，補完システムと英語圏の人間（以後，nativeと呼ぶ）との両方に与える．このnativeも，試行者Aと同様に補完システムの動作原理を知らされていない．次に，補完システムが生成した出力テキストとnativeが作成したテキストの両方を，もとの試行者Aが自分の意図に合う内容にするべくポスト・エディットする．ポスト・エディット量に関して，システムの出力テキストに対するポスト・エディッ

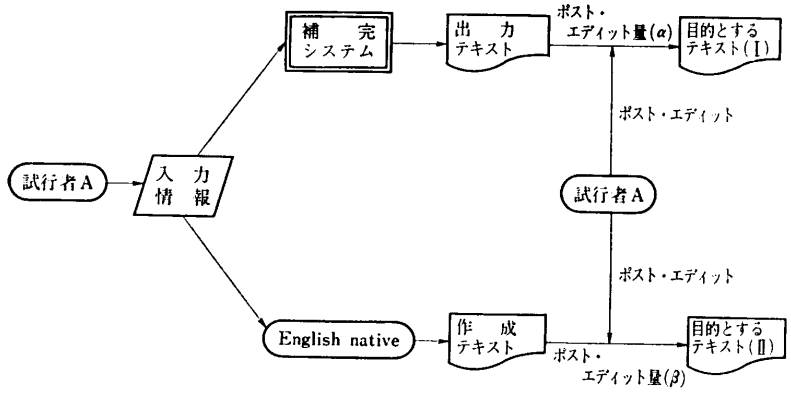


図 7 評価Ⅱの実験系
Fig. 7 System for the evaluation-Ⅱ.

ト量を α, native の作成テキストへのポスト・エディット量を β とおく．ポスト・エディット量は，次の規則に従って算定する．

補完システムが対象とするBasic English は，意味のかたまりとしてのチャンクによって英文を構成していくという指針をもつ．補完システムの解析部は，その指針に基づいて設計されている．したがって，チャンクに対応する構文上の単位である文，句，語を同等に扱った．ゆえに，ポスト・エディットにおいてもチャンクを一単位として扱う．すなわち，

```
WENT OUT MY HOME.  AT STATION.
TOOK A TRAIN IN TIME.
TO TOKYO AND TOOK A MEAL WITH A GUIDE WHO WAS MY FRIEND.
```

(a) 入力情報

```
I WENT OUT TO MAKE SOMETHING.
I WENT TO MAKE SOMETHING BY SOME WAY TO STATION.
I TOOK A TRAIN AT STATION IN TIME.
I TOOK TO TOKYO AT STATION, AND I TOOK A MEAL WITH A GUIDE
   WHO WAS MY FRIEND WHILE I WAS AT STATION.
```

(b) システムの出力テキスト

```
I WENT OUT TO THE STATION.
AT THE STATION I COULD TAKE A TRAIN IN TIME.
I MADE FOR TOKYO AND TOOK A MEAL WITH A GUIDE WHO WAS MY FRIEND.
```

(c) native の作成テキスト

```
I WENT OUT MY HOME.
    (置換：TO MAKE SOMETHING →MY HOME )
△.
    (削除：I WENT TO MAKE SOMETHING BY SOME WAY TO STATION.)
I TOOK A TRAIN AT THE STATION IN TIME.
    (挿入：THE )
I MADE FOR TOKYO △ ,AND I TOOK A MEAL WITH A GUIDE WHO WAS MY FRIEND△ .
    (置換：TOOK TO →MADE FOR, 削除：AT STATION,
     削除：WHILE I WAS AT STATION)
```

(d) ポスト・エディットされた出力テキスト

図 8 ポスト・エディットの例
Fig. 8 Example of post-editing.

1) 一単位の挿入は補完能力の明白な欠如の結果なのでカウントは1とする．

2) 一単位の削除は不必要な情報の補完の結果なのでカウントは0とする．

3) 一単位の置換は削除と挿入の組合せと考えてカウントは1とする．

例えば，補完システムの出力テキストに対して試行者Aが，3回の挿入，2回の削除，1回の置換を施した場合は，$\alpha=4$ である．冗長な情報は補完能力の欠如を示すものではなく，また削除するのは考え出す作業よりも楽であるから，利用者の注意を喚起する目的でなるべく出力するようにしてある．この算定法は，一見したところ字面を対象にしているように見えるが，意味のかたまりをカウントしているという点で補完システムの補完能力の評価としての意味をもつ．図8に，入力情報，補完システムの出力テキスト，native の作成テキスト，ポスト・エディット後のテキストの例を示す．例におけるポスト・エディット量は，$\alpha=3$, $\beta=1$ である．

ポスト・エディット量 α, β に対して，

$$R_H=\frac{1+\ln(\beta+1)}{1+\ln(\alpha+1)} \qquad (1)$$

を**補完比**と呼び，人間との補完能力の比較，もしくは補完システム間の補完能力の比較の目安とする．

$\alpha, \beta \geq 0$

であるので，

$\alpha+1>0$, $\beta+1>0$

として対数式の真数が0になるのを防いでいる．さらに，

$1+\ln(\alpha+1)\geq 1$

だから，(1)式の分母は0にならない．分子の形は分母に合わせた．ポスト・エディット量の対数をとった理由は，人間の主観や感覚が物理量の対数に比例するとする Weber/Fechner の法則[5]に従ったことによる．

補完比 R_H は，次のような意味をもつ．

i) $0<R_H<1$…システムの補完能力が人間よりも劣る．R_H が0に近いほど，システムの補完能力は低い．

表 3 評価 II の結果
Table 3 Result of the evaluation-II.

	補完比	補正評価値（点）
最大値	1.0	99.3
平均値	0.58	35.8
最小値	0.32	0

ii）$R_H=1$…システムと人間の補完能力は同等である．

iii）$1<R_H$…システムの補完能力が人間よりも優れている．$R_H<\infty$ であるので，そのレンジで R_H の解釈を考慮する．

ところで，一般に入力情報が少ないほど補完処理に負荷がかかり，結果的にポスト・エディット量は増大する．たとえ入力情報中の語数を一定に統制したとしても，その意味内容まで均一に統制することは実際上不可能である．そこで，補完システムと native の2系統にまったく同一の情報を与えてテキストを生成させ，それらのポスト・エディット量の比をとることによって入力情報の多少が出力側のポスト・エディット量に影響を与えないよう考慮した．

4.2 評価IIの結果

評価Iで用いた 40 題の不完全テキストを入力情報とし，同一の native（カナダ人，女性，20歳）を系に固定して評価を行った．補完比を求めた結果を表3に示す．補完比の平均値は 0.58 であり，$\beta=0$, $\alpha=1\sim2$ の例がほとんどであった．α が大きい場合の原因は，少ない入力情報に対し過大なポスト・エディットを施したような例が多くみられた．ここで α が大きいということは，システムによって補完してほしい情報が補完されなかったということを示しているのであって，意味を成さない出力テキストであったというのとは異なる．

4.3 補正評価値

補完比は，複数の補完比間の比較を目的とした指標であって，単独値の解釈によってシステムの補完能力を評価するには不向きである．そこで，補完比に補正を加えて単独値でも有益であるように拡張した．

すなわち，英語教育の経験者が native の作成したテキストを正答（100 点）として，それに対するシステムの出力テキストに 0 点～100 点の範囲で主観採点値 P を与える．P と補完比との間には相関があることが予想される．そこで，Weber/Fechner の法則に基づき，

$$\ln\left(100\cdot\frac{\beta+1}{\alpha+1}\right)$$

と P との相関を調べたところ有意（危険率 5%）であったので，

$$P=k\cdot\ln\left(100\cdot\frac{\beta+1}{\alpha+1}\right)+b$$

とおいて回帰的に k, b を求め，次式を得た．

$$P=57.83\ln\left(100\cdot\frac{\beta+1}{\alpha+1}\right)-167$$

$$=57.83\ln\frac{\beta+1}{\alpha+1}+99.32.$$

ただし，$(\beta+1)/(\alpha+1)<0.18$ の場合，$P<0$ となってしまうので結局，次式を実際に用いる．

$$\begin{cases} P=57.83\ln\dfrac{\beta+1}{\alpha+1}+99.32 & \left(0.18<\dfrac{\beta+1}{\alpha+1}\right) \\ P=0 & \left(0<\dfrac{\beta+1}{\alpha+1}\leq 0.18\right) \end{cases} \quad (2)$$

(2)式によって求められる P の値は，システムの出力テキストを仮に採点した場合，何点ぐらいの得点をとるような能力を持つかの指標として用いることができる．この P を評価 II の**補正評価値**と呼ぶことにする．表3には，補正評価値が補完比とともに示されている．主観採点値は5点刻みで与えられたので，補正評価値もそれほど高い精度は期待できない．実際に与えられた主観採点値は，図4の例文の場合で 80 点，図5の例文の場合で 50 点であった．

5. 考　　察

（1） 評価Iによって，内容的に質の良いテキストを生成する場合に強力に寄与しているのが連想型・推論型であることがわかった．この結果から，省略現象の一部，および補完体系の性質が明らかになった．

（2） 評価IIの手法はかなり有効であり，この種のシステムの能力を数量的かつ1元的に評価する一つの方法を示せた．ただし，ポスト・エディット量の算定の仕方が粗すぎるのではないかという指摘があり，例えば文・句・語のレベルでテキスト全体への各々の貢献度が違うのならば，それを考慮したグレードを導入するなどが考えられる．

（3） 現在の補完体系が，はたして省略現象全体をカバーしているかという点が問題になった．典型的な場面における行動系列を含むようなテキストの補完結果の多くが，質の良くないテキスト群に分類されたからである．意味構造上，スクリプト[6]を用いて表現さ

れ得るようなテキストの補完結果が，意外にも低品質であった．この結果をふまえて，日本語テキストにおける省略表現の復元システム内に MOPs モデル[7]を導入し，好結果を得た[8),9)]．その詳細は別の機会にゆずる．

（4） 定形文書のような，表層レベルにおける絞切り型表現を含むテキストを扱う能力が欠如している*．これには，解析と生成の両面において特別な機構を必要とする．すなわち，解析時における定形パターンの認識と特殊な言いまわしの解析，そして生成時における定形パターンの発生と適切な慣用語句の選択である．現行の補完システムでは，深層構造に変換された時点で定形文書特有の表現を示す情報を失うので，不適切な補完や無意味な出力テキストを生成することが多い．

6. むすび

補完システムの利用者が頭の中に描いた内容と補完システムが物理的に生成したテキストの内容との比較評価を行うために，実験心理学的な手法を導入した．システムの出力テキストの内容は，入力情報に対する一つの解釈であり，この解釈が利用者のあらかじめ意図していた内容にどれだけ近いかを評価するのが，本論文における評価の目的である．この評価により，補完体系の性質を明らかにした．さらに，質の良い内容のテキストをより多く生成させるための，補完システムに対する tuning 技法に有効なことが確かめられている[4)]．また，本研究で用いた評価実験系は実験心理学的にも初めての試みであり，その有効性から将来の自然言語理解システムの評価において，いずれ必要になる手法であると思われる．

謝辞 心理学的見地から有益な助言をいただいた京都大学文学部心理学教室の乾敏郎助手，および評価実験用に補完システムを拡張してくれた大学院学生堂坂浩二君に感謝の意を表します．また本研究の一部は文部省科学研究費「特定研究」によった．

* この点に関して，京都大学工学部の辻井潤一助教授よりご指摘があった．

参考文献

1) 松永，小川，田中：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，情報処理学会「知識工学と人工知能」研究会資料，28-2 (1982).
2) 唐沢，田村，松永：英文補完生成システムの補完能力評価，情報処理学会第27回全国大会予稿集，4D-3, pp. 1139-1140 (1983).
3) 唐沢，小川，田村：英文補完システムの補完能力評価，情報処理学会「自然言語処理」研究会資料，43-4 (1984).
4) 唐沢，小川，田村：英文テキスト補完システムの tuning 技法，情報処理学会第29回全国大会予稿集，6N-3, pp. 1271-1272 (1984).
5) Rumelhart, D. E.: *An Introduction to Human Information Processing*, John Wiley and Sons, Inc., New York (1977).
6) Schank, R. C. and Abelson, R. P.: *Scripts, Plans, Goals and Understanding*, Lawrence Erlbaum Associates, Inc., New Jersey (1977).
7) Schank, R. C.: Reminding and Memory Organization: An Introduction to MOPs, Yael Univ. Research Report #170 (1979).
8) 唐沢，堂坂，小川，田村：省略を含む日本語テキストの復元に関するコンピュータ・モデル，日本認知科学会第1回大会発表論文集，B-2, pp. 32-33 (1984).
9) 堂坂，唐沢，小川，田村：Prolog による省略された日本語の復元システム，情報処理学会第29回会国大会予稿集，2N-2, pp. 1187-1188 (1984).
本論文投稿後，本研究と同様の考え方に基づく文献 10)「翻訳システムの評価」が発表されたので参考にされたい．
10) 長尾，辻井：Mu プロジェクトにおける日英翻訳結果の評価，情報処理学会「自然言語処理」研究会資料，47-11 (1985).

(昭和59年10月5日受付)
(昭和60年4月25日採録)